Unit 7

Point estimation

Introduction

In earlier units an important distinction has been drawn between *sample* quantities on the one hand – values calculated from data, such as the sample mean and the sample standard deviation – and corresponding *population* quantities on the other. The latter arise when a statistical model is assumed to be an adequate representation of the underlying variation in the population. Usually the model family is specified (binomial, Poisson, normal, ...), but the indexing parameters (the binomial probability p, the Poisson parameter λ, the normal mean μ and normal variance σ^2, and so on) might be unknown; indeed, they usually will be unknown. Often one of the main reasons for collecting data is to *estimate*, from a sample, the value(s) of the model parameter(s).

Many different sample quantities can be used to estimate population parameters. For example, if there are x successes in a sample of n independent Bernoulli trials, each with probability p of success, then the proportion of successes in the sample, x/n, can be used to estimate p, the unknown parameter of the binomial distribution. Similarly, for normal distributions, the sample mean $\overline{x}$ can be used as an estimate of the population mean μ, and the sample variance s^2 as an estimate of the population variance σ^2.

Sometimes, however, it is not totally clear which sample quantity should be used to estimate a population parameter. For instance, should the sample mean or the sample variance be used to estimate the Poisson parameter λ, as λ is both the population mean and the population variance of this distribution? Similarly, for a normal population, μ is the population median as well as the population mean, so should the sample mean or the sample median be used as an estimate of μ?

In this unit, the focus is on *point estimation*: the task of providing a single value to estimate a population parameter. To determine a point estimate, or just *estimate* for short, an estimating formula, or *estimator*, is applied to the data available. For example, the formula for a sample mean – an estimator – might be applied to a sample to obtain the numerical value of the sample mean – an estimate – which is a quantity that might be used to estimate the population mean – a population parameter.

Point estimation contrasts with interval estimation, or the providing of an interval of values to estimate a parameter; interval estimation is the subject of Unit 8.

However, different samples can produce different estimates for the same parameter using the same estimator. For example, when using the sample mean as an estimator of a population mean, you already know that the observed value of the sample mean varies from sample to sample. This means that the sample mean will not, in general, equal the population mean. (Even if one of the various sample means happens to equal the population mean, the others almost certainly won't. What is more, we wouldn't know which of the sample means is equal to the population mean.) It is, however, desirable to obtain estimators that generally have values which are as close to the true value of the parameter of interest as possible. In Section 1, a number of examples are examined in which data have been collected on a random variable and a population parameter is to

be estimated from the data. Attractive properties of an estimator are also considered and illustrated with examples.

In Section 2, a very important approach to parameter estimation, with broad application, is introduced. This approach is the *method of maximum likelihood*, or just *maximum likelihood* for short. For any given probability distribution, maximum likelihood is a general method for obtaining an estimator of the parameter of the distribution. Differentiation is the usual technique used in determining a maximum likelihood estimator, so in Subsection 3.1 we review the results on differentiation that we need here. The results of Subsection 3.1 are applied in Subsection 3.2 to obtain maximum likelihood estimates of parameters from samples of data. More generally, maximum likelihood estimators are derived in Subsection 4.1, and some properties of the method of maximum likelihood are given in Subsection 4.2.

1 Principles of point estimation

When some representative statistical model has been proposed for the variation observed in a random variable, **point estimation** is the process of using the data available to estimate the unknown value of the parameter (or parameters) of the model. The single number obtained from the data is a **(point) estimate** of the parameter. In Subsection 1.1, examples of **(point) estimators** – formulas which deliver (point) estimates when applied to the data – are given; and in Subsection 1.2, the question of 'What makes a good estimator?' is addressed. In Subsection 1.3, you will use your computer to explore and compare some estimators.

1.1 Point estimators

We will start by considering an example of an estimator.

Example 1 *Counts of the leech* Helobdella

In a research study, 103 water samples were collected from a lake. To avoid confusion with our different use of the word 'sample' everywhere else, let us call these water samples 'volumes'. The number of specimens of the leech *Helobdella* contained in each volume was counted. More than half of the volumes collected (58 of them) were free of this contamination, but all the other volumes contained at least one leech, and three contained five or more leeches. Table 1 gives the frequencies of the different counts.

Table 1 Counts of the leech *Helobdella* in 103 water volumes

Count	0	1	2	3	4	5	6	7	8	≥ 9
Frequency	58	25	13	2	2	1	1	0	1	0

(Source: Jeffers, J.N.R. (1978) *An Introduction to Systems Analysis with Ecological Applications*, London, Edward Arnold)

A plausible model for the observed variation in the counts is a Poisson distribution. Since the parameter λ of a Poisson distribution is the mean of the distribution, a natural estimate of λ is the sample mean $\overline{x}$. In this case, the sample mean is

$$\overline{x} = \frac{0 \times 58 + 1 \times 25 + 2 \times 13 + \cdots + 8 \times 1}{58 + 25 + 13 + \cdots + 1} = \frac{84}{103} \simeq 0.816.$$

So 0.816 is a point estimate of the unknown Poisson mean λ.

The dataset was presented in frequency form in Table 1 and the sample mean, $\overline{x}$, was calculated in the usual way using those frequencies. Those frequencies were, in turn, obtained from the raw data $x_1, x_2, \ldots, x_{103}$, where the xs denote the number of *Helobdella* leeches in each of the 103 water volumes. In fact,

$$x_1 = 2,\ x_2 = 0,\ x_3 = 1,\ x_4 = 4,\ x_5 = 0,\ \ldots,\ x_{103} = 0.$$

In terms of the raw data, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_{103}}{103} = \frac{2 + 0 + \cdots + 0}{103} = \frac{84}{103} \simeq 0.816,$$

as before. The observed values that comprise the raw data can be thought of as a particular collection of 103 independent observations $X_1, X_2, \ldots, X_{103}$ on the random variable $X \sim \text{Poisson}(\lambda)$. Let

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_{103}}{103}.$$

Then $\overline{X}$ is a random variable whose observed value (0.816) is our estimate of λ; $\overline{X}$ is an estimator of λ.

As has been mentioned before, it is often useful, as here, to distinguish between random variables, by denoting them by upper-case letters such as X, and their observed, sample, values, by denoting them by the corresponding lower-case letters, in this case x.

The freshwater leech *Helobdella robusta*

In traditional printing using movable type letters, the letters were stored in a 'type case', with small letters in the 'lower case' and capital letters in the 'upper case'

In Example 1, a procedure or *estimating formula* was used which may be expressed as follows. Collect a total of n water volumes and count the numbers of leeches $X_1, X_2, \ldots, X_n$ in the volumes; find the total number of leeches $X_1 + X_2 + \cdots + X_n$, and divide this number by n to obtain the average number of leeches in a volume of water. In other words, the formula used to estimate the parameter λ of the population model Poisson(λ) is the random variable

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

It is called an *estimator* of λ. With different datasets from the same situation, different values of the estimator would be obtained; these are *estimates* of λ. For example, for the particular dataset in Example 1, the estimate takes the value $84/103 \simeq 0.816$.

Example 1 illustrates the following essential features of a point estimation problem.

- Some *data*: if we have no data, then we cannot make an estimate.
- A *probability model* for the way the data were generated. In Example 1, $X \sim \text{Poisson}(\lambda)$.
- The model involves a parameter whose value is unknown: this is the value we wish to estimate. In Example 1, the parameter is λ.
- An estimating formula or *estimator* of the parameter: this formula is obtained from the model (and includes symbols for the data rather than their numerical values). In Example 1, the estimator is $\overline{X} = (X_1 + X_2 + \cdots + X_{103})/103$.
- The value of the estimator given by entering the data into the estimating formula, that is, the *estimate* for the parameter. In Example 1, the estimate is $\overline{x} = (x_1 + x_2 + \cdots + x_{103})/103 \simeq 0.816$.

We will now introduce an important piece of notation.

Strictly speaking, a 'hat' symbol is called a 'caret' symbol. No, not this sort of carat ...

... nor this sort of carrot

Hat notation

It is common practice to use a 'hat' symbol to indicate an estimate of a parameter. So estimates of μ and p (say) are denoted by $\widehat{\mu}$ and $\widehat{p}$ (pronounced 'mew-hat' and 'p-hat'), respectively.

The hat notation is also used for an estimator. The estimator of μ in Example 1 is $\widehat{\mu} = \overline{X} = (X_1 + X_2 + \cdots + X_n)/n$. It would be unwieldy to develop a separate notation to distinguish estimates and estimators, so this is not done.

Examples 2 and 3 further illustrate the features of a point estimation problem.

Example 2 *Alveolar–bronchiolar adenomas in mice*

An adenoma is a benign tumour originating in a gland.

In a research experiment into the incidence of alveolar–bronchiolar (respiratory) adenomas in mice, several historical datasets on groups of mice were examined. (Source: Tamura, R.N. and Young, S.S. (1987) 'A stabilised moment estimator for the beta-binomial distribution', *Biometrics*, vol. 43, no. 4, pp. 813–24.) One of the groups contained 54 mice. After examination, six of the 54 mice were found to have adenomas. These are the data from the first group. Assuming independence between mice, the experiment consists of observing an outcome x on a binomial random variable X, which represents the number of mice in the sample who have adenomas. The probability model is $X \sim B(n, p)$ where n is the sample size and p is the unknown parameter that we wish to estimate; p is the probability that a mouse has an adenoma. The obvious estimator (estimating formula) for p is the proportion of mice in the sample who have adenomas,

$$\widehat{p} = X/n.$$

For this first group, the number observed was $x = 6$ and the sample size was $n = 54$, so the estimate of p is $\widehat{p} = 6/54 = 1/9$, or about 0.111.

A different group might have involved a different number, n, of subjects but used the same experimental design. Making the same assumptions, our estimating procedure would again be: observe the value of the random variable X and divide this observed value by n. That is, X/n is again the estimator of p although, obviously, its value will generally differ in different experiments.

The group size n varies, but we know its value for any group, so it is not a *random* variable. It is an observed (or perhaps chosen) value, so it is denoted by a lower-case letter.

Indeed, the experiment involved altogether 23 groups of mice. Examination of the other 22 groups resulted in different estimates for the proportion of affected mice in the wider population, from as low as $0/20 = 0$ in one sample of 20, through a variety of different values such as $4/47 \simeq 0.085$, to as high as $4/20 = 0.2$.

Notice that the 23 different groups of mice (samples) are assumed to be similar so that the observed proportions of alveolar–bronchiolar adenomas in each group can all be viewed as different estimates of the overall proportion of alveolar–bronchiolar adenomas in an appropriate underlying population of mice.

Example 3 concerns the normal distribution. This differs from Examples 1 and 2 in that the distribution is continuous, rather than discrete, and involves two unknown parameters rather than just one. Nevertheless, the key features of the estimation problem are the same.

Example 3 *Chest measurements of nineteenth-century Scottish soldiers*

The chest circumference of each of 5732 nineteenth-century Scottish soldiers was measured (in inches, to the nearest inch). These measurements are the data. A histogram of the data suggests that they might be observations from a normal distribution. Hence the probability model has the form $X \sim N(\mu, \sigma^2)$, where X denotes the chest circumference in inches of a nineteenth-century Scottish soldier. The data are observations on the random variables $X_1, X_2, \ldots, X_{5732}$, and the unknown parameters are μ and σ^2. The obvious estimators of these quantities are the sample mean

The data are given in Table 2 of Unit 6 and a histogram of the data in Figure 2 of Unit 6.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

where $n = 5732$. Computing the sample mean and variance for the observed data values gives $\overline{x} \simeq 39.8489$ inches and $s^2 \simeq 4.2989$ inches^2, so these are the estimates of μ and σ^2.

Later in Unit 6, the normal distribution with $\mu = 40$ and $\sigma^2 = 4$ was used as a simplified version of this model for these data.

1.2 What makes a good estimator?

An estimator is a random variable, so it has a probability distribution. This distribution is called the **sampling distribution of the estimator**.

We investigated the sampling distribution of the sample mean and sample total in Section 6 of Unit 6.

Looking at the mean and variance of its sampling distribution can give a good idea of how well the estimator in question can be expected to perform.

Activity 1 *Mean and variance of a Poisson sample mean*

Example 1 concerned a research study about the number of leeches of the genus *Helobdella* that were found in each of 103 water volumes from a lake. This study resulted in an estimate of the Poisson parameter λ of $\overline{x} = 0.816$, which is an observation on the random variable $\overline{X}$. Now suppose that a similar study was carried out under similar circumstances, but that only 48 water volumes were collected. Denote the number of leeches in these volumes by $y_1, y_2, \ldots, y_{48}$. Assuming the same Poisson model, if we again use the sample mean to estimate the unknown parameter λ, our estimate from this second study would be

It is much clearer to use different symbols for values from a different study.

$$\overline{y} = \frac{y_1 + y_2 + \cdots + y_{48}}{48}.$$

This is an observation on the random variable

$$\overline{Y} = \frac{Y_1 + Y_2 + \cdots + Y_{48}}{48},$$

our estimator of λ in this case. Notice that both $\overline{X}$ and $\overline{Y}$ are estimators of the same parameter λ.

What is the variance of a Poisson distribution whose mean is λ? Write down in terms of the unknown parameter λ the mean and variance of the random variable $\overline{X}$ and of the random variable $\overline{Y}$. Hint: recall from Unit 6 that for any random sample taken from a population with mean μ and variance σ^2, the mean and variance of the sample mean $\overline{Z}$, say, are given by $E(\overline{Z}) = \mu$ and $V(\overline{Z}) = \sigma^2/n$, respectively, where n is the sample size.

So what, in particular, does the mean of an estimator tell us about the performance of the estimator? Some estimators may tend to give estimates that are too high on average, while others may tend to give estimates that are too low on average. (In science and engineering, the behaviour of the average of an estimator is termed its 'accuracy'.) Clearly, an estimator that avoids being 'wrong' on average is desirable. Such estimators are said to be unbiased. More specifically, an estimator is said to be *unbiased* if its expected value is equal to the parameter being estimated, that is, if an estimator gives estimates that are on average equal to the parameter being estimated. Unbiasedness is a theoretical property of estimators under an assumed model.

Unbiased estimators

Suppose that $\widehat{\theta}$ is an estimator of a parameter θ. Then $\widehat{\theta}$ is an **unbiased** estimator of θ if

$$E(\widehat{\theta}) = \theta.$$

The unknown parameter in estimation problems is often denoted θ, which is the Greek lower-case letter theta, pronounced 'theta'.

Example 4 *Unbiasedness of the sample mean*

Recall from Subsection 6.1 of Unit 6 that whatever the underlying population distribution, if $\overline{X}$ is the sample mean and μ is the population mean, then

$$E(\overline{X}) = \mu.$$

We can therefore formalise this aspect of the notion that the sample mean is a good estimator of the population mean: the sample mean is an unbiased estimator of the population mean.

So in the case of a Poisson(λ) distribution, for instance, the sample mean is, on these grounds, a reasonable choice of estimator of the Poisson parameter, λ, since λ is the population mean in this case.

The result found in Example 4 is important and worth highlighting:

The sample mean is an unbiased estimator of the population mean.

Activity 2 *An unbiased estimator of the binomial parameter*

Suppose that the random variable X follows a binomial distribution $B(n, p)$. Show that $\widehat{p} = X/n$ is an unbiased estimator of p.

Bias of estimators

Suppose again that $\widehat{\theta}$ is an estimator of a parameter θ. An estimator $\widehat{\theta}$ is **biased** if it is not unbiased. In this case, the **bias** of $\widehat{\theta}$ is

$$E(\widehat{\theta}) - \theta.$$

An estimator is said to be **positively biased** if

$$E(\widehat{\theta}) > \theta;$$

such an estimator gives estimates that are too high on average. Similarly, an estimator is said to be **negatively biased** if

$$E(\widehat{\theta}) < \theta;$$

such an estimator gives estimates that are too low on average.

The one that got away ... bias in estimation of fish size?

Example 5 *A biased estimator of the geometric parameter*

Suppose that a random variable X follows the geometric distribution with parameter p. From Unit 4, the mean of the geometric distribution is $\mu = 1/p$. We know, therefore, from Example 4, that X itself, being the

sample mean of this sample of size $n = 1$, is an unbiased estimator of the population mean μ.

However, the above concerns estimation of the quantity $1/p$, not of the parameter p itself. No problem, one might say – use the estimator $1/X$ to estimate the parameter p. And yes, this is a reasonable thing to do. But is $\widehat{p} = 1/X$ an unbiased estimator of p? The answer turns out to be no. In fact, $E(\widehat{p})$ is a complicated function of p that it is far beyond the scope of this module to derive. However, this complicated function of p is certainly not equal to p itself, so $\widehat{p} = 1/X$ is a biased estimator of p.

It is generally true that $E(1/X) \neq 1/E(X)$. In this case, if you really want to know, $E(1/X) = -p \log p/(1-p)$, and $\widehat{p} = 1/X$ is positively biased.

Unbiasedness is not the be all and end all of desired properties of a good estimator. It is also preferable for an estimator to have *low variance*. Then, estimates resulting from statistical experiments can be expected to almost always be quite close to the parameter that is being estimated when the bias is also low. (In science and engineering, this property is that of high 'precision'.) In Activity 1, for example, you found the mean and variance of the estimators $\widehat{\lambda}_1 = \overline{X}$ and $\widehat{\lambda}_2 = \overline{Y}$ of the Poisson parameter λ based on two different samples of data from the same situation. Both $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ are unbiased estimators of λ as their means are λ, but $\widehat{\lambda}_1$ has the smaller variance because it is based on a larger sample. So, in that sense, $\widehat{\lambda}_1$ is a better estimator of λ than is $\widehat{\lambda}_2$.

It is not unreasonable to sometimes use a biased estimator (for example, like the one in Example 5), provided that both its bias is small and its variance is low.

Other desirable qualities of an estimator relate to its properties as the sample size, n, increases. The performance of an estimator often improves as the sample size increases: its variance usually decreases, and its bias, if any, may well decrease also. One might expect this kind of performance on the basis that the more data you have, the better you should be able to estimate the parameter(s) of interest.

Activity 3 *The mean of a normal distribution*

Suppose that a random sample of twelve observations is taken from the normal distribution, $N(\mu, 25)$.

(a) Is the sample mean an unbiased estimator of μ? What is its variance?

(b) Would the variance of the sample mean decrease if the sample size were increased?

In any given estimation problem, there is not always one clear estimator to use: there may be several possible alternative estimators that could be employed. The question that naturally arises is: 'Which estimator is likely to lead to "better" estimates?' This is illustrated in Example 6 and Activity 4.

Example 6 *Observations with different variances*

Suppose that independent observations X_1, X_2 and X_3 come from normal distributions that have the same mean, μ, but whose variances are 1, 4 and 9. That is,

$$X_1 \sim N(\mu, 1), \quad X_2 \sim N(\mu, 4) \quad \text{and} \quad X_3 \sim N(\mu, 9).$$

One possible estimator of μ is

$$\widehat{\mu}_1 = \tfrac{1}{3}(X_1 + X_2 + X_3).$$

Then

$$\begin{aligned} E(\widehat{\mu}_1) &= E\{\tfrac{1}{3}(X_1 + X_2 + X_3)\} \\ &= \tfrac{1}{3}E(X_1 + X_2 + X_3) \\ &= \tfrac{1}{3}\{E(X_1) + E(X_2) + E(X_3)\} \\ &= \tfrac{1}{3}(\mu + \mu + \mu) = \mu. \end{aligned}$$

Here, we have first used $E(aX + b) = a\,E(X) + b$ with $a = \frac{1}{3}$, $b = 0$ and then $E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n)$ with $n = 3$; both results are initially from Unit 4. We have shown that the expected value of the estimator $\widehat{\mu}_1$ is simply the unknown parameter μ. So $\widehat{\mu}_1$ is an unbiased estimator of μ.

However, let us consider another possible estimator, namely

$$\widehat{\mu}_2 = \tfrac{1}{49}(36X_1 + 9X_2 + 4X_3).$$

This estimator gives greater weight (more importance) to X_1 than to the other variables, which seems appropriate as X_1 has a smaller variance than the other variables and, in consequence, is likely to be closer to μ than the other variables. Similarly, $\widehat{\mu}_2$ gives less weight to X_3 than to the other variables, which also seems appropriate, as X_3 has the largest variance. We have that

$$\begin{aligned} E(\widehat{\mu}_2) &= E\{\tfrac{1}{49}(36X_1 + 9X_2 + 4X_3)\} \\ &= \tfrac{1}{49}E(36X_1 + 9X_2 + 4X_3). \end{aligned}$$

But the expectation term equals $E(36X_1) + E(9X_2) + E(4X_3)$ by applying $E(Y_1 + Y_2 + Y_3)$, where $Y_1 = 36X_1$, $Y_2 = 9X_2$ and $Y_3 = 4X_3$. So

In general, $E(a_1X_1 + a_2X_2 + a_3X_3) = a_1E(X_1) + a_2E(X_2) + a_3E(X_3)$.

$$\begin{aligned} E(\widehat{\mu}_2) &= \tfrac{1}{49}\{E(36X_1) + E(9X_2) + E(4X_3)\} \\ &= \tfrac{1}{49}\{36E(X_1) + 9E(X_2) + 4E(X_3)\} \\ &= \tfrac{1}{49}(36\mu + 9\mu + 4\mu) = \mu. \end{aligned}$$

Hence $\widehat{\mu}_2$ is also an unbiased estimator of μ.

Thus both $\widehat{\mu}_1$ and $\widehat{\mu}_2$ have the desirable property of being unbiased estimators of μ. However, the values that they take are unlikely to be identical. For example, if the data values are $x_1 = 5.2$, $x_2 = 4.7$ and $x_3 = 4.8$, then

$$\widehat{\mu}_1 = \tfrac{1}{3}(5.2 + 4.7 + 4.8) = 4.9$$

while

$$\widehat{\mu}_2 = \tfrac{1}{49}(36 \times 5.2 + 9 \times 4.7 + 4 \times 4.8) \simeq 5.08.$$

Should we prefer $\widehat{\mu}_1$ or $\widehat{\mu}_2$ as an estimator of μ?

As noted earlier, it is preferable if an estimator has a small variance as well as being unbiased. Hence we should examine the variances of $\widehat{\mu}_1$ and $\widehat{\mu}_2$.

Now,

$$\begin{aligned} V(\widehat{\mu}_1) &= V\{\tfrac{1}{3}(X_1 + X_2 + X_3)\} \\ &= \left(\tfrac{1}{3}\right)^2 V(X_1 + X_2 + X_3) \\ &= \tfrac{1}{9}\{V(X_1) + V(X_2) + V(X_3)\} \\ &= \tfrac{1}{9}(1 + 4 + 9) = \tfrac{14}{9} \simeq 1.56. \end{aligned}$$

This time, we have first used $V(aX + b) = a^2\,V(X)$ with $a = \frac{1}{3}$ and then $V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n)$ with $n = 3$ since X_1, X_2 and X_3 are independent; again, results are initially from Unit 4.

For $\widehat{\mu}_2$, we have

$$\begin{aligned} V(\widehat{\mu}_2) &= V\{\tfrac{1}{49}(36X_1 + 9X_2 + 4X_3)\} \\ &= \left(\tfrac{1}{49}\right)^2 V(36X_1 + 9X_2 + 4X_3) \\ &= \left(\tfrac{1}{49}\right)^2 \{V(36X_1) + V(9X_2) + V(4X_3)\} \\ &= \left(\tfrac{1}{49}\right)^2 \{36^2\,V(X_1) + 9^2\,V(X_2) + 4^2\,V(X_3)\} \\ &= \tfrac{1}{2401}(1296 \times 1 + 81 \times 4 + 16 \times 9) = \tfrac{1764}{2401} \simeq 0.73. \end{aligned}$$

The third line follows from the second because $V(Y_1 + Y_2 + Y_3) = V(Y_1) + V(Y_2) + V(Y_3)$, where $Y_1 = 36X_1$, $Y_2 = 9X_2$, $Y_3 = 4X_3$ (and the Ys are independent).

Thus the variance of $\widehat{\mu}_2$ is much smaller than the variance of $\widehat{\mu}_1$, so $\widehat{\mu}_2$ is the better estimator.

In forming $\widehat{\mu}_2$ in Example 6, the coefficients of X_1, X_2 and X_3 were made proportional to 36, 9 and 4 because

$$36\,V(X_1) = 9\,V(X_2) = 4\,V(X_3),$$

all being equal to 36. It can be shown, but will not be so here, that this choice means that $\widehat{\mu}_2$ has a smaller variance than any other unbiased estimator of μ that is linear in X_1, X_2 and X_3. You can, however, check this claim against another possible estimator that is linear in X_1, X_2 and X_3 in the following activity.

Activity 4 *Another estimator for observations with different variances*

For X_1, X_2 and X_3 as defined in Example 6, consider the estimator

$$\widehat{\mu}_3 = \tfrac{1}{11}(6X_1 + 3X_2 + 2X_3).$$

(These coefficients satisfy $6\,S(X_1) = 3\,S(X_2) = 2\,S(X_3) = 6$, where S denotes standard deviation.)

(a) Show that $\widehat{\mu}_3$ is an unbiased estimator of μ.

(b) Calculate $V(\widehat{\mu}_3)$, and hence verify that $V(\widehat{\mu}_3)$ is greater than $V(\widehat{\mu}_2)$.

(c) Which of $\widehat{\mu}_2$ and $\widehat{\mu}_3$ is the better estimator of μ?

The examples in this subsection have shown that it can sometimes be straightforward to determine whether an estimator is unbiased and to calculate its variance. These qualities can be used to choose between alternative estimators.

To finish the subsection, we give an example in which obtaining an unbiased estimator involves somewhat trickier mathematics. You will not be expected to reproduce all the algebraic details, and can ignore those in the accompanying screencast if you wish. However, the results of the example are important, as they relate to the task of obtaining an unbiased estimator of a population variance.

How sharp were prehistoric spear heads? A case study in point estimation?

Example 7 *An unbiased estimator of the population variance*

By definition, if μ and σ^2 are the population mean and the population variance, then

$$\sigma^2 = E[(X - \mu)^2].$$

Suppose we want to estimate σ^2 from a random sample of n observations, $X_1, X_2, \ldots, X_n$. We estimate $\mu = E(X)$ by the sample mean $\overline{X} = \sum_{i=1}^{n} X_i/n$ so, by analogy, an obvious estimator of σ^2 is

$$W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The non-standard notation W will help to keep things clearer later.

However, the usual estimator of a population variance is the sample variance,

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

To understand why this latter estimator is generally preferred, we will first determine the expected value of $\sum_{i=1}^{n}(X_i - \overline{X})^2$, which occurs in both estimators. From this, we will be able to conclude that S^2 is an unbiased estimator of σ^2 while W is a biased estimator, which is the reason that S^2 is the preferable estimator in most situations.

Screencast 7.1 goes through the mathematical manipulations involved in showing that

$$E\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = (n-1)\sigma^2. \qquad (1)$$

As already indicated, this screencast is optional.

Screencast 7.1 Verifying Equation (1) (optional)

It follows from Equation (1) that

$$E(S^2) = E\left\{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = \frac{1}{n-1}E\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}$$
$$= \frac{1}{n-1} \times (n-1)\,\sigma^2 = \sigma^2.$$

So S^2 is an unbiased estimator of σ^2, as stated above.

Activity 5 *A biased estimator of the population variance*

(a) Use Equation (1) to show that

$$W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

is a biased estimator of σ^2.

(b) Calculate the bias of W, and interpret what the bias tells us.

1.3 Exploring and comparing estimators by computer

The work in this subsection consists of a chapter of Computer Book B. You will use computer animations to compare the properties of different estimators of particular parameters.

Refer to Chapter 3 of Computer Book B for the work in this subsection.

Exercise on Section 1

Exercise 1 *Weight gains of pigs*

From prior experience, it is known that the weight gain of pigs on a high-protein diet has a standard deviation of 9.5 kg; on a low-protein diet the standard deviation is 8.2 kg. A new supplement is added to the diets. With this supplement, let a pig's average weight gain be μ_1 on the high-protein diet and μ_2 on the low-protein diet. To estimate $\mu_1 - \mu_2$, the difference in average weight gain on the two diets, eleven pigs are put on the high-protein diet and ten pigs are put on the low-protein diet; both diets include the supplement. The supplement does not change the standard deviations of weight gain.

A natural estimator of $\mu_1 - \mu_2$ is $\overline{X}_1 - \overline{X}_2$, where $\overline{X}_1$ is the mean weight gain of the eleven pigs on the high-protein diet and $\overline{X}_2$ is the mean weight gain of the ten pigs on the low-protein diet. The estimators $\overline{X}_1$ and $\overline{X}_2$ are independent because they are based on different pigs.

(a) Show that $\overline{X}_1 - \overline{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$.

(b) Find the variance of the estimator $\overline{X}_1 - \overline{X}_2$.

(c) If one further pig were available for the experiment, what would be the variance of the estimator if the pig were put on the high-protein diet?

What would be the variance of the estimator if the pig were put on the low-protein diet? Which of the two diets should the pig be put on in order to get the best estimate of $\mu_1 - \mu_2$? (In either case, the difference between sample means remains an unbiased estimator of the difference between population means.)

2 The method of maximum likelihood

In the previous section, estimates and estimators were introduced and we discussed what makes a good estimator. In this section, we will introduce one of the most widely used methods for finding estimates and estimators.

In our day-to-day lives we regularly guess the most likely explanation for things that happen. If somebody walks past your window holding an open umbrella, the most likely reason is that it is raining. If you press a light switch and nothing happens, you might conclude that 'the light bulb has probably gone', because that seems the most likely reason for the failure. If many people are waiting at a bus stop, you might conclude that a bus should come soon because it seems likely that one has not come for some time.

... or it might be snowing

The approach of asking 'What could best explain this?' or 'What is most likely?' can be used to estimate the unknown parameter(s) of a probability model. Given a set of observations, we can ask: 'What value of the parameter is most likely to have given rise to these observations?' To this end, we will define a *likelihood function* which encapsulates the probability of the observations arising from the model for each of the possible values of its parameter. And then we will choose our estimate of the parameter to be the value which maximises this likelihood. This value is referred to as the *maximum likelihood estimate* of the parameter. The whole methodology of forming estimators in this way is called **maximum likelihood estimation**. It is arguably the most important method of constructing estimators: it is highly versatile – it can be used in an enormous variety of situations – and the estimators that it yields have good properties.

See Subsection 2.1 for a detailed explanation.

The basic idea underlying maximum likelihood estimation is introduced in detail in Subsection 2.1. Partly for notational convenience, discrete and continuous distributions are considered separately; but the underlying principle is essentially the same for both forms of distribution. Maximum likelihood estimation for discrete distributions is discussed in Subsection 2.1 and for continuous distributions in Subsection 2.2. To introduce ideas, in this section, we use graphs to obtain approximate values of maximum likelihood estimates, and summarise how far we have got to in Subsection 2.3; calculus will then be used in Sections 3 and 4 to obtain estimates and estimators exactly.

2.1 Discrete probability models

In this subsection, it is assumed that observations are collected on a discrete random variable X. The random variable X therefore has a probability mass function $p(x) = P(X = x)$. Suppose that the variation in observations on X is to be modelled by a probability distribution indexed by a single unknown parameter θ. To emphasise that there is a parameter θ involved, the probability mass function may be written $p(x; \theta)$. So

$$P(X = x) = p(x; \theta)$$

for all values x in the range of X.

Suppose initially that our sample consists of just a single observation x from the discrete probability model with p.m.f. $p(x; \theta)$. This means that, before collecting the sample, the probability of observing the sample value x is

$$p(x; \theta). \tag{2}$$

But this probability depends on the value of θ; if $\theta = \theta_1$, say, then $p(x; \theta_1)$ takes one value, while if $\theta = \theta_2$, say, then $p(x; \theta_2)$ takes a different value. After collecting the sample, we know the value of x but we still don't know the value of θ. So instead of our usual habit of thinking of $p(x; \theta)$ as a function of x (for fixed but unknown θ), we can now think of $p(x; \theta)$ as a function of θ (for fixed and known x). It is this function of θ that we will denote by

$$L(\theta) = p(x; \theta) \tag{3}$$

and call the **likelihood of θ based on a single observation** from a discrete model. Here, and in more general situations, we usually abbreviate this terminology to just the **likelihood function** or often just the **likelihood**.

Having obtained the likelihood of the data value – its probability of arising given our model for each value of its parameter θ – we can choose our estimate of θ to maximise this likelihood function, that is, as the value of θ which makes the data value we observed the most probable to have arisen under the model. Said again, we choose the **maximum likelihood estimate**, $\widehat{\theta}$, of θ as the value of θ which maximises the likelihood function $L(\theta)$.

Let's see how this works out for an observation from the binomial distribution.

Example 8 *Rolling a biased die*

Suppose a die is rolled ten times and on seven of these rolls it lands showing a 5. If θ is the probability that the die shows a 5 when rolled once, then it seems likely that θ is much greater than $\frac{1}{6}$ and that the die is biased.

What is the value of θ that is most likely to give seven 5s in ten rolls? Assuming the rolls of the die are independent, the number of 5s has a binomial distribution, $B(10, \theta)$. Hence the probability of observing exactly

seven 5s in ten rolls is

$$p(7;\theta) = \binom{10}{7}\theta^7(1-\theta)^3 = \frac{10!}{7!\,3!}\theta^7(1-\theta)^3 = 120\,\theta^7(1-\theta)^3.$$

Now, instead of thinking of $p(7;\theta)$ as the probability of the sample value we observed (i.e. 7), given a fixed (if unknown) value for θ, we turn things round and consider $p(7;\theta)$ as the likelihood function for θ given the known value, 7, of the observation:

$$L(\theta) = p(7;\theta) = 120\,\theta^7(1-\theta)^3.$$

The method of maximum likelihood involves finding the value of θ that makes this likelihood as large as possible.

Now, in the binomial model, the parameter θ can take any value between 0 and 1. The likelihood, being a function of θ, is therefore a function of a continuous argument, θ, even though the data value (and model) with which we are dealing is discrete. Table 2 gives the value of $L(\theta)$ for some of the possible values of θ; Figure 1 plots $L(\theta)$ as a function of θ over its entire range $(0, 1)$.

Table 2 The value of the likelihood for various values of θ

θ	0	0.2	0.4	0.6	0.7	0.8	1
$L(\theta)$	0	0.0008	0.0425	0.2150	0.2668	0.2013	0

The table suggests that $L(\theta)$ increases from 0 at $\theta = 0$ to a peak at about $\theta = 0.7$ and then decreases. This is confirmed by Figure 1, in which $L(\theta)$ is plotted against θ. The figure shows that the likelihood is maximised when θ is approximately 0.7. So $\widehat{\theta}$, the maximum likelihood estimate of θ, is approximately 0.7. (In fact, $\frac{7}{10} = 0.7$ is the exact value of the maximum likelihood estimate.)

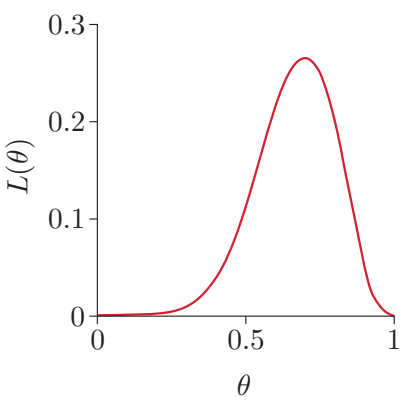

Figure 1 A graph of $L(\theta)$

You will learn how to find the exact value of a maximum likelihood estimate in the next section.

Activity 6 *Likelihood for a binomial parameter*

Example 2 concerned adenomas in mice. In one group of 54 mice, six had adenomas. Let θ be the (unknown) proportion of mice in the whole population that have adenomas.

(a) Write down $L(\theta)$, the likelihood for θ (based on the above data).

(b) Evaluate $L(\theta)$ at $\theta = 0.11$, and $\theta = 0.12$ and hence complete the following table.

Table 3

θ	0.09	0.10	0.11	0.12	0.13
$L(\theta)$	0.1484	0.1643			0.1558

(c) Use the values in the completed table to sketch a graph with $L(\theta)$ plotted against θ for values of θ between 0.09 and 0.13. (Even though $L(\theta)$ is calculated for only five values of θ in part (b), remember that θ

can actually take any value between 0 and 1; in this case, $L(\theta)$ turns out to be very small for values of θ less than 0.09 or greater than 0.13.)

(d) Find an approximate value for the maximum likelihood estimate of θ.

So far, so much fuss about nothing, you might think! We have gone through all this likelihood rigmarole in Example 8 and Activity 6 just to get 'obvious' estimates of the probability θ in the binomial case. Well, yes, the estimates might be obvious so far, but the approach remains useful in far more complicated situations when the appropriate value of an estimate is nothing like so obvious. And it is an encouraging property of the general maximum likelihood approach that it does reduce to simple, 'obvious' estimates in simple cases.

So, to continue to develop the maximum likelihood methodology, let us now consider the more usual situation where, for the purposes of estimating the value of θ, a random sample $X_1, X_2, \ldots, X_n$ of size n, where n is greater than 1, is collected. Under the model with p.m.f. $p(x;\theta)$, the probability that X_1 takes the value x_1, say, is $p(x_1;\theta)$; the probability that X_2 equals x_2 is $p(x_2;\theta)$; and so on. Although the term 'random sample' has been used a number of times in the module already, we will now make explicit what has largely been implicit before: the observations in a random sample are assumed to be independent of one another. It follows that the probability that $X_1 = x_1$ *and* $X_2 = x_2$ is the product of the probability that $X_1 = x_1$ and the probability that $X_2 = x_2$:

$$P(X_1 = x_1, X_2 = x_2) = p(x_1;\theta) \times p(x_2;\theta).$$

And, more generally, it follows that the probability that $X_1 = x_1$ *and* $X_2 = x_2$ *and* $\cdots$ *and* $X_n = x_n$ – that is, the probability of obtaining $x_1, x_2, \ldots, x_n$ as the collection of sample values – is the product of all the individual probabilities:

$$p(x_1;\theta) \times p(x_2;\theta) \times \cdots \times p(x_n;\theta). \tag{4}$$

Now, this expression gives the probability that our actual sample arose, given the true, but unknown, value of θ. As such, it is the direct extension to $n > 1$ of the probability that our sample arose when $n = 1$ given in Expression (2).

So, using Expression (4) in place of Expression (2), the argument proceeds just as it did before. First, as we do not know θ, we cannot be sure what the true value of this probability is. However, we can work out the value of this probability for various guessed values of θ. In doing this, we are treating the probability as a function of θ: we switch things round from considering Expression (4) as a function of $x_1, x_2, \ldots, x_n$ for fixed, if unknown, θ to a function of θ for fixed values of $x_1, x_2, \ldots, x_n$ provided by the sample. For each particular value of θ, this function tells us how likely we are to obtain our particular sample. So it seems reasonable to estimate θ as the value that gives maximum probability to the particular sample that actually arose; that is, we should choose $\widehat{\theta}$ to maximise the quantity in Expression (4).

The probability in Expression (4) is called the likelihood of θ for the sample $x_1, x_2, \ldots, x_n$ or, usually, simply the **likelihood**, and is denoted by $L(\theta)$:

$$L(\theta) = p(x_1; \theta) \times p(x_2; \theta) \times \cdots \times p(x_n; \theta). \quad (5)$$

This likelihood, really the **likelihood function**, is to be considered as a function of the unknown parameter θ, and is the extension of Equation (3) to the situation where $n > 1$. Note that the possible values of θ usually lie in some continuous interval regardless of whether the data are discrete or continuous.

The method of maximum likelihood involves answering the question: 'What value of θ maximises the chance of observing the random sample that was, in fact, obtained?' So we define the *maximum likelihood estimate* of θ, denoted by $\widehat{\theta}$, as the value of θ that maximises the likelihood $L(\theta)$ given by Equation (5). A ubiquitous abbreviation in statistics is the one for the maximum likelihood estimate: *MLE*.

The method of maximum likelihood in the discrete case is summarised in the following box.

The method of maximum likelihood for discrete data

If X is a discrete random variable with probability mass function $p(x; \theta)$, where θ is an unknown parameter, then the likelihood for the random sample $x_1, x_2, \ldots, x_n$ is denoted by $L(\theta)$ and is given by

$$L(\theta) = p(x_1; \theta) \times p(x_2; \theta) \times \cdots \times p(x_n; \theta).$$

The method of maximum likelihood involves finding the value $\widehat{\theta}$ of θ that maximises the likelihood $L(\theta)$. This value is the maximum likelihood estimate (MLE) of θ.

MLE. So famous, there's even a film about it. Or perhaps there it's short for 'My Little Eye'.

In the next example and the following activities, the method of maximum likelihood is used with datasets consisting of more than one observation.

Example 9 *Estimating the parameter of a geometric distribution*

Consider a very small artificial dataset of three observations, $x_1 = 3$, $x_2 = 4$ and $x_3 = 8$. Suppose that these are independent observations from a geometric distribution with unknown parameter. For consistency with the current development, the parameter indexing the geometric distribution will temporarily be referred to as θ. (Conventionally, it is denoted by p.)

The probability mass function for the geometric distribution with parameter θ is

$$p(x; \theta) = (1 - \theta)^{x-1}\theta, \quad x = 1, 2, 3, \ldots.$$

In forming likelihoods, use is often made of the result that $q^a q^b \dots q^k = q^{a+b+\cdots+k}$, where q is any quantity.

It follows that the likelihood for this particular random sample of size 3 is given by

$$\begin{aligned} L(\theta) &= p(x_1;\theta) \times p(x_2;\theta) \times p(x_3;\theta) \\ &= (1-\theta)^{x_1-1}\theta \times (1-\theta)^{x_2-1}\theta \times (1-\theta)^{x_3-1}\theta \\ &= (1-\theta)^{3-1}\theta \times (1-\theta)^{4-1}\theta \times (1-\theta)^{8-1}\theta \\ &= (1-\theta)^{2+3+7}\theta^3 \\ &= (1-\theta)^{12}\theta^3. \end{aligned}$$

The likelihood $L(\theta)$ is a function of the unknown parameter θ which lies somewhere between 0 and 1: for different values of θ, the function will take different values. We need to find the value of θ at which this function is maximised. Table 4 gives values of the likelihood $L(\theta)$ for various values of θ; $L(\theta)$ is graphed as a function of all values of θ in $(0, 1)$ in Figure 2.

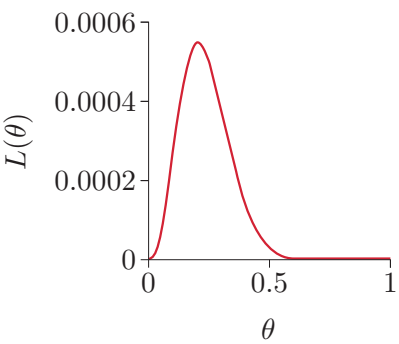

Figure 2 A graph of $L(\theta)$

Table 4 The values of the likelihood for various values of θ

θ	0	0.2	0.4	0.6	0.8	1
$L(\theta)$	0	0.00054976	0.00013931	0.00000362	0.00000000	0

These calculations suggest that the likelihood is maximised somewhere between $\theta = 0$ and $\theta = 0.4$ – possibly at $\theta = 0.2$ itself. As you can see from the graph of the likelihood in Figure 2, the likelihood is maximised when the value of θ is approximately 0.2. (In fact, 0.2 is the exact value of $\widehat{\theta}$.)

Activity 7 *Different data from a geometric model*

Consider another very small artificial dataset that can be assumed to come from a geometric distribution, with a different value of its parameter θ (which we wish to estimate). This dataset consists of the four independent observations $x_1 = 1$, $x_2 = 2$, $x_3 = 1$ and $x_4 = 3$.

(a) Show that the likelihood for this particular random sample of size 4 is given by

$$L(\theta) = (1-\theta)^3\theta^4.$$

(b) A graph of the likelihood obtained in part (a) is shown in Figure 3, for all θ in $(0, 1)$. Use this figure to find the approximate value of the MLE of θ.

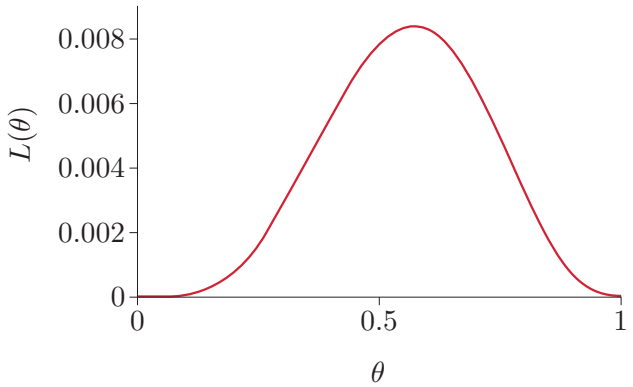

Figure 3 A graph of $L(\theta)$

The next activity is different from the previous examples and activities in this subsection, in that the probability model being applied (and which is indexed by a single unknown parameter) is not one of the standard families. However, the principle is the same: a value of the parameter is sought that maximises the likelihood for the sample that was actually observed.

Activity 8 *The leaves of Indian creeper plants*

The leaves of the Indian creeper plant *Pharbitis nil* can be variegated or unvariegated and, at the same time, faded or unfaded. The resulting four leaf types are denoted 0 for unvariegated and unfaded, v for variegated and unfaded, f for unvariegated and faded, and vf for variegated and faded.

In an experiment, plants were crossed. Of 290 offspring plants observed, the four types of leaf occurred with frequencies which are given in Table 5.

A theory allowing for a phenomenon called 'genetic linkage' assumes that the observations in this experiment might have arisen from a probability distribution indexed by an unknown parameter θ. According to this theory, the different types of leaf have the probabilities given in Table 5. Since all these probabilities must be non-negative, something we do know about θ is that it must lie between $-\frac{1}{16}$ and $\frac{3}{16}$. (These limits are because we must have $\frac{1}{16} + \theta > 0$ and $\frac{3}{16} - \theta > 0$; in this model, θ is not itself a probability, so might be negative.)

A *Pharbitis nil* hybrid with variegated leaves

Table 5 *Pharbitis nil* model probabilities and observed frequencies

Type of leaf	Unvariegated and unfaded (0)	Variegated and unfaded (v)	Unvariegated and faded (f)	Variegated and faded (vf)
Probability	$\frac{9}{16} + \theta$	$\frac{3}{16} - \theta$	$\frac{3}{16} - \theta$	$\frac{1}{16} + \theta$
Frequency	187	37	35	31

(Source: Bailey, N.T.J. (1961) *Introduction to the Mathematical Theory of Genetic Linkage*, Oxford, Clarendon Press, p. 41)

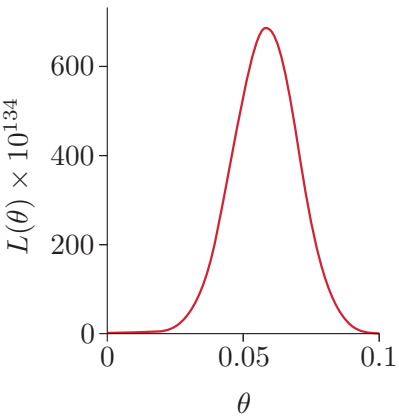

Figure 4 A graph of $L(\theta) \times 10^{134}$ for $0 < \theta < 0.1$

(a) Write down the likelihood $L(\theta)$ associated with this experiment.

(b) The likelihood function turns out to take extremely small values, so, to make it visible, Figure 4 shows $L(\theta) \times 10^{134}$. Even when rescaled like this, the likelihood function takes essentially zero values outside the range $0 < \theta < 0.1$, so $L(\theta)$ is plotted only on this range. What is the approximate MLE of θ?

2.2 Continuous probability models

So far, attention has been restricted to discrete probability models. In this subsection, the maximum likelihood approach is developed for continuous random variables.

What is the likelihood of the unknown parameter θ for a random sample $x_1, x_2, \ldots, x_n$ from a continuous distribution? In the discrete case, we were able to say that the likelihood is the product of the probabilities $p(x_i; \theta)$. However, for continuous random variables, the probability of obtaining any particular value, such as x_1, is effectively zero, so the probability of obtaining any particular sample of values, such as $x_1, x_2, \ldots, x_n$, is effectively zero, also. The key to moving from the discrete case to the continuous case is to replace the probability mass function of the discrete case with the probability density function in the continuous case. So in the continuous case, the likelihood is obtained by replacing the p.m.f. $p(x; \theta)$ throughout Equation (5) by the p.d.f. $f(x; \theta)$. Thus, in the continuous case, the likelihood may be written as

Here, $f(x)$ has been rewritten as $f(x; \theta)$ to emphasise the dependence of f on θ.

$$L(\theta) = f(x_1; \theta) \times f(x_2; \theta) \times \cdots \times f(x_n; \theta). \tag{6}$$

The notation $L(\theta)$ again expresses the fact that the likelihood is thought of as a function of θ (not of the fixed values $x_1, x_2, \ldots, x_n$). The method of maximum likelihood involves finding the value $\widehat{\theta}$ of θ that maximises this likelihood.

In this way, instead of choosing $\widehat{\theta}$ to maximise the probability of the observed data, as in the discrete case, we are actually choosing $\widehat{\theta}$ to maximise the density of the observed data in the continuous case.

Unsuccessful flood defence

Example 10 *Estimating the exponential parameter*

Flood protection was built around a river in a small town to reduce the risk of flooding. However, it was not very successful. After its construction, the town first flooded five years later, the second flood was eight years after that, and the third flood came after a further seven years. Suppose that the time between floods is an observation from an exponential distribution with parameter $\theta > 0$, and that the MLE of θ is required. For an exponential distribution with parameter θ, the probability density function is

$$f(x; \theta) = \theta e^{-\theta x}, \quad x \geq 0.$$

Thus, for the random sample $x_1 = 5$, $x_2 = 8$, $x_3 = 7$ from this distribution, the likelihood of θ is

$$
\begin{aligned}
L(\theta) &= f(x_1;\theta) \times f(x_2;\theta) \times f(x_3;\theta) \\
&= \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \times \theta e^{-\theta x_3} \\
&= \theta e^{-\theta 5} \times \theta e^{-\theta 8} \times \theta e^{-\theta 7} \\
&= \theta^3 e^{-\theta(5+8+7)} = \theta^3 e^{-20\theta}.
\end{aligned}
$$

As with discrete probability models, the MLE of θ can be determined approximately from a graph of $L(\theta)$ against θ. Again because all values of the likelihood are very small, Figure 5 shows the likelihood multiplied by a large factor, this time $10^4 = 10\,000$, and because even this rescaled likelihood is still very small for $\theta \geq 0.5$, it is plotted only for $\theta < 0.5$ (recall that θ can actually take any positive value). Figure 5 shows that the MLE of θ is about 0.15. (In fact, 0.15 is the exact value of $\widehat{\theta}$.)

This illustrative example and Activity 9 to follow assume that the risk of flooding is constant over time. This assumption might be invalid because of, for example, climate change.

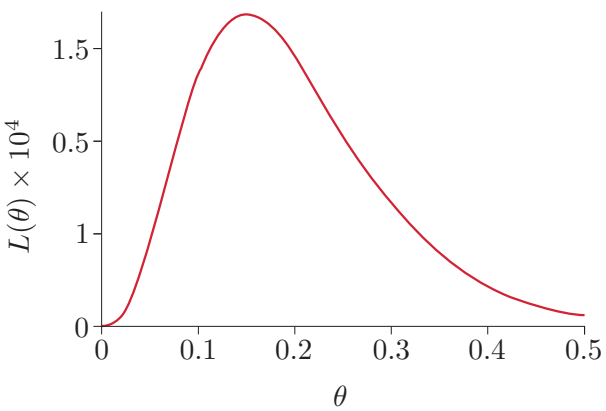

Figure 5 A graph of $L(\theta) \times 10^4$ for $0 < \theta < 0.5$

It is quite typical that likelihoods for continuous data take very small values. To be able to work with simpler numbers, in Figure 5 (as in Figure 4 in the discrete case), the likelihood was rescaled by multiplication by a large factor. The amount of this rescaling is arbitrary, and unimportant in the sense that the maximiser of the likelihood function is in the same place regardless of what (positive) factor the likelihood is multiplied by. Also, even after rescaling, the likelihood often remains very small for many of the possible values of θ, so in Figures 4 and 5, $L(\theta)$ was plotted only over that interval of values of θ for which $L(\theta)$ is reasonably large. We will continue to employ this type of rescaling and plotting over a limited interval of values of θ where appropriate in further figures in this unit.

Activity 9 *Estimating another exponential parameter*

In another small town with a similar geography and potential for flooding, flood protection built many years ago seems to be rather more effective. In subsequent years, the town has flooded just twice: once 25 years after construction, then again some 31 years after that. Suppose that here too the times between floods are observations from an exponential distribution, but with a different value for its parameter $\theta > 0$, and that the MLE of θ is required.

These flood defences are holding, protecting properties to the right

(a) What is the likelihood for θ based on this history of flooding?

(b) Figure 6 shows (a rescaled version of) $L(\theta)$ for these data. What is the approximate MLE of θ?

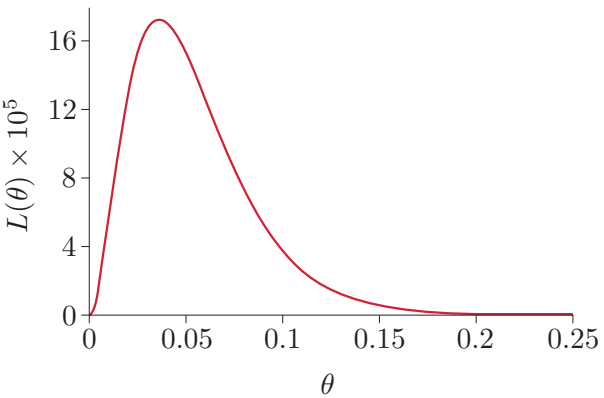

Figure 6 A graph of $L(\theta) \times 10^5$ for $0 < \theta < 0.25$

Activity 10 *The Rayleigh distribution*

Let X denote wind speed. This is a continuous, positive, random variable that is sometimes modelled as following the *Rayleigh distribution*. The probability density function of the Rayleigh distribution is given by

$$f(x;\theta) = \frac{x}{\theta^2}\, e^{-x^2/2\theta^2}, \quad x > 0,$$

with $\theta > 0$. (This distribution has various further applications, including a natural role in MRI (magnetic resonance imaging) scanning.)

(a) Suppose that 22.2, 2.8, 4.0, 13.9, 11.7 and 8.3 are a random sample of six observations of X taken at the site of a possible wind farm on different days, measured in km/h. Show that the likelihood $L(\theta)$ of θ is given (approximately) by

$$L(\theta) = \frac{335\,621}{\theta^{12}}\, e^{-457.8/\theta^2}.$$

(b) Evaluate $L(\theta)$ at $\theta = 8.50$ and $\theta = 9.00$, and hence complete the following table giving $L(\theta)$ for various values of θ.

Table 6

θ	8.25	8.50	8.75	9.00	9.25
$L(\theta) \times 10^9$	4.048		4.216		4.059

(c) Use the values in the table to sketch a plot of $L(\theta) \times 10^9$ against θ for values of θ between 8.25 and 9.25. Hence find an approximate value for $\widehat{\theta}$.

2.3 The story so far

Let us start this subsection with some more notation. In the same way that $\sum$ is used to denote a sum, for example,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n,$$

so $\prod$ is used to denote a product, for example,

$$\prod_{i=1}^{n} x_i = x_1 \times x_2 \times \cdots \times x_n.$$

Π is the Greek upper-case letter Pi

This allows us to write the likelihood in a slightly more compact way, and to remind you of the workings of maximum likelihood estimation for both discrete and continuous data.

The method of maximum likelihood

If X is a random variable with a distribution with unknown parameter θ, then the likelihood of θ for the random sample $x_1, x_2, \ldots, x_n$, or likelihood for short, is denoted by $L(\theta)$ and is given by

$$L(\theta) = \begin{cases} \prod_{i=1}^{n} p(x_i; \theta) & \text{if } X \text{ is discrete} \\ \prod_{i=1}^{n} f(x_i; \theta) & \text{if } X \text{ is continuous,} \end{cases} \tag{7}$$

where $p(x; \theta)$ is the probability mass function and $f(x; \theta)$ is the probability density function.

The method of maximum likelihood involves finding the value $\widehat{\theta}$ of θ that maximises the likelihood $L(\theta)$. This value is the maximum likelihood estimate (MLE) of θ.

The logo of a shoe shop in Seattle, USA

Let us now make some further remarks about the likelihood. (The second one is a reinforcement of something you have seen already.)

- 'Likelihood' has the above specific, technical, meaning to statisticians. This meaning is not the same as its everyday use as a synonym for 'chance' or even 'probability'.
- 'The likelihood' is shorthand for 'the likelihood function'. The likelihood is a function of θ. This represents a turnaround given that its constituent parts, probability mass or density functions, are usually thought of as functions of 'x', indexed by θ. In the likelihood, θ becomes the argument of the function while 'the xs', that is, $x_1, x_2, \ldots, x_n$, are thought of as fixed quantities.
- The likelihood is positive: $L(\theta) > 0$. This is because all its constituent parts, which are multiplied together, are positive. (In the discrete case, the probabilities of observed events must be positive. But even in the continuous case, the density f *at an observed value* x_i cannot be zero.) If your likelihood is zero or negative for any value of θ, you've made a mistake in its construction!

If you are not confident with how to formulate a likelihood function or are not clear on its interpretation, Screencast 7.2 might be helpful to you.

Screencast 7.2 Formulating and understanding a likelihood

In this section, we have made plots of $L(\theta)$, as a function of θ. These plots have then been used to determine (approximately) the point at which $L(\theta)$ is maximised: this point gives (approximately) the maximum likelihood estimate $\widehat{\theta}$. Exact maximisation methods are, however, really needed. These could be numerical or mathematical. Numerical methods comprise computational algorithms that hone in, to very high levels of precision, on the maximum of the likelihood function. Numerical methods come into their own when models become much more complicated and parameters much more numerous. We won't investigate any of them in M248. When likelihood functions are relatively simple and the number of parameters is small – particularly, when the number of parameters is 1! – their maxima can usually be found mathematically, using calculus: the remainder of this unit focuses on using calculus to find MLEs.

The man is presumed not to have enough practice to improve noticeably as he goes along!

Exercise on Section 2

Exercise 2 *Clay pigeon shooting*

Clay pigeon shooting, also known as clay target shooting, consists of individuals shooting guns at clay targets that are fired into the air by a machine. A man tried this out for the first time, stopping after he had hit four targets. He took three shots to first hit a target, one shot to hit the next target, and two shots each to hit two further targets. Suppose that the number of shots he takes to hit a clay target can be modelled by a geometric distribution with parameter θ.

(a) Given these data, find an expression for $L(\theta)$, the likelihood of θ.

(b) Evaluate $L(\theta)$ at $\theta = 0.4$ and $\theta = 0.5$, and hence complete the following table.

Table 7

θ	0.3	0.4	0.5	0.6	0.7
$L(\theta)$	0.0019			0.0033	0.0019

(c) Use the values in the completed table to draw a rough graph with $L(\theta)$ plotted against θ for values of θ between 0.3 and 0.7.

(d) Use your graph to find an approximate value for the MLE of θ.

3 Using calculus to find maximum likelihood estimates

In Section 2, graphs were used to determine maximum likelihood estimates. More commonly, calculus is used because it yields exact values. Also, a graph can be used to determine the estimate only for a particular set of data, whereas calculus can be used to derive the formula for a maximum likelihood estimator, the general formula that gives the maximum likelihood estimate when the observed data values are entered into it. In this section, however, we will continue to confine attention to maximum likelihood estimates themselves.

The key to finding maxima is differentiation. In Subsection 3.1, we revise results on differentiation that are required in this unit. In Subsection 3.2, we will use the results to find some maximum likelihood estimates.

These results should already be familiar to you.

3.1 Differentiation of powers, polynomials and exponentials, and their combinations and products

In this unit, we will need to differentiate quantities like the power $2x^3$, the polynomial

$$4 + 3x + x^2 - 5x^3 + 2x^7,$$

and the exponential function e^{-3x}, and functions made up by combining these quantities in particular ways.

You *integrated* powers and polynomials in Subsection 3.1 of Unit 2.

Differentiating powers, polynomials and exponentials

Let us start by differentiating powers.

The derivative of a constant times a power

If $f(x) = ax^k$, then the derivative of $f(x)$ is

$$\frac{d}{dx}f(x) = f'(x) = kax^{k-1}.$$

In words, we multiply by the power of x and then reduce the power of x by 1. Notice that the multiplicative constant a remains unchanged.

Two important special cases of this are the derivatives of a constant and of x itself:

- since $a = ax^0$, it follows that $f'(x) = 0$ when $f(x) = a$
- since $ax = ax^1$, it follows that $f'(x) = a$ when $f(x) = ax$.

Both notations for the derivative, $\frac{d}{dx}f(x)$ and $f'(x)$, will be used in what follows.

Example 11 *Differentiating powers*

To illustrate applying this rule:

if $f(x) = 4$, then $f'(x) = 0$;

if $f(x) = 3x$, then $f'(x) = 3$;

if $f(x) = 4x^3$, then $f'(x) = 3 \times 4x^{3-1} = 12x^2$;

if $f(x) = \dfrac{5}{x^2} = 5x^{-2}$, then $f'(x) = -2 \times 5x^{-2-1} = -10x^{-3} = -\dfrac{10}{x^3}$;

and if $f(x) = 2\sqrt{x} = 2x^{1/2}$, then $f'(x) = \dfrac{1}{2} \times 2x^{(1/2)-1} = x^{-1/2} = \dfrac{1}{\sqrt{x}}$.

A political map of the world: using colours to differentiate powers?

Activity 11 *Differentiating powers*

Find the derivatives of each of the following functions.

(a) $6x^2$ (b) $4x^{5.1}$ (c) $\dfrac{5}{x^4}$ (d) $-4x\sqrt{x}$ (e) $27x$ (f) $-\dfrac{3}{\sqrt{x}}$

Now we can extend from powers to polynomials; the key is the method for dealing with a sum of functions.

The derivative of a sum of functions

Suppose that $f(x) = g(x) + h(x) + \cdots + q(x)$, where $g(x)$, $h(x)$, …, $q(x)$ are any functions of x. Then

$$f'(x) = g'(x) + h'(x) + \cdots + q'(x).$$

That is, the derivative of a sum is the sum of the derivatives.

More generally, if $a, b, \ldots, k$ are constants and $f(x) = a\,g(x) + b\,h(x) + \cdots + k\,q(x)$, then

$$f'(x) = a\,g'(x) + b\,h'(x) + \cdots + k\,q'(x).$$

Example 12 *Differentiating a polynomial*

As a first example, if

$$f(x) = 4x^3 + \frac{5}{x^2} + 3x,$$

then

$$f'(x) = \frac{d}{dx}(4x^3) + \frac{d}{dx}\left(\frac{5}{x^2}\right) + \frac{d}{dx}(3x) = 12x^2 - \frac{10}{x^3} + 3.$$

Here, we have used the derivatives of the individual power terms already found in Example 11.

Activity 12 *Differentiating polynomials*

Find the derivatives of the following.

(a) $4 + 3x + x^2 - 5x^3 + 2x^7$

(b) $4 - \dfrac{3}{\sqrt{x}} - \dfrac{2}{x^3}$

The exponential function arises in, for example, the p.m.f. of the Poisson distribution and the p.d.f. of the exponential distribution, so we have to be able to differentiate it as well. Happily, this is no more difficult than differentiating powers of x.

The derivative of a constant times an exponential function

If $f(x) = ae^{kx}$, then the derivative of $f(x)$ is

$$f'(x) = kae^{kx}.$$

In words, we just multiply by the coefficient of x in the exponential function.

An important special case of this is the derivative of the exponential function itself. Since e^x corresponds to $a = k = 1$, it follows that $f'(x) = e^x$ when $f(x) = e^x$.

Example 13 *Differentiating exponentials*

To illustrate applying this rule:

if $f(x) = 2e^{3x}$, then $f'(x) = 3 \times 2e^{3x} = 6e^{3x}$;

if $f(x) = e^{-x}$, then $f'(x) = (-1) \times e^{-x} = -e^{-x}$;

if $f(x) = 3e^{x/3}$, then $f'(x) = \dfrac{1}{3} \times 3e^{x/3} = e^{x/3}$.

The second of these examples, the derivative of e^{-x}, which results in *minus* e^{-x}, is particularly useful in statistics.

Activity 13 *Differentiating exponentials*

Find the derivatives of each of the following functions.

(a) $6e^{x/2}$ (b) $3e^{-3x}$ (c) $10e^{-0.1x}$

If you are unsure about the basic differentiation methods that you have just worked through, Screencast 7.3 might be of assistance.

Screencast 7.3 Differentiating a polynomial plus an exponential

Differentiating functions of functions, and products

You will also need to use the rules for differentiating certain functions of functions and, since the likelihood is a product of functions, for differentiating products of functions. Let us start with functions of functions.

Suppose that you need to differentiate the function

$$f(x) = (2x + 1)^8.$$

You couldn't make such an expansion if the power were, say, 8.5.

This is a power of a polynomial. Now, you *could*, in this case, expand the power, thereby writing $f(x)$ as a long polynomial. But it turns out to be much easier to treat $f(x)$ as a function – the power 8 – of another function – the polynomial $2x + 1$. The result for differentiating a function of a function is known as the *chain rule*.

The chain rule

Suppose that the function $f(x)$ can be written in terms of other functions g and h as

$$f(x) = h(g(x)). \tag{8}$$

Then

$$f'(x) = g'(x)\, h'(g(x)).$$

We will use the chain rule particularly in the case of powers of polynomials.

Example 14 *Differentiating powers of polynomials*

To illustrate applying this rule, consider differentiating

$$f(x) = (2x + 1)^8.$$

This is of the form of Equation (8) if we set

$$h(y) = y^8 \quad \text{and} \quad y = g(x) = 2x + 1.$$

Now,

$$h'(y) = 8y^7 \quad \text{and} \quad g'(x) = 2.$$

It follows that

$$f'(x) = 2 \times 8y^7 = 2 \times 8(2x + 1)^7 = 16(2x + 1)^7.$$

Activity 14 *Differentiating powers of polynomials*

Find the derivatives of each of the following functions.

(a) $(1-x)^k$ (b) $12\left(1+2x+\dfrac{2}{\sqrt{x}}\right)^4$

And finally, what about a product of functions?

This will be particularly important in what follows because likelihoods are products of simpler functions.

The derivative of a product of two functions

Suppose that $f(x) = g(x) \times h(x)$, where $g(x)$ and $h(x)$ are any functions of x. Then

$$f'(x) = g'(x)\,h(x) + g(x)\,h'(x). \tag{9}$$

Example 15 *Differentiating a product*

Suppose that

$$f(x) = 3x(2x+1)^8.$$

Then $f(x)$ can be written as $g(x)\,h(x)$ where

$$g(x) = 3x \quad \text{and} \quad h(x) = (2x+1)^8.$$

It doesn't matter which function you call g and which you call h.

Now,

$$g'(x) = 3 \quad \text{and, from Example 14,} \quad h'(x) = 16(2x+1)^7.$$

It follows from Equation (9) that

$$f'(x) = 3 \times (2x+1)^8 + 3x \times 16(2x+1)^7.$$

We would then normally simplify the expression for $f'(x)$ by extracting common factors, so that

Extracting common factors in this way will be particularly helpful when finding MLEs.

$$f'(x) = 3(2x+1)^7(2x+1+16x) = 3(2x+1)^7(18x+1).$$

Activity 15 *Differentiating products*

Find the derivatives of each of the following functions.

(a) $x^2(1-x)^3$ (b) xe^{-x}

In marketing, product differentiation is the process of making your product stand out from the others

The chain rule and the rule for differentiating a product of two functions are put to good use in an example in Screencast 7.4.

Screencast 7.4 Differentiating a power of a polynomial times an exponential

3.2 Maximum likelihood estimates

In Example 8, we formed the likelihood for fitting a binomial distribution with parameter θ to particular observations on the rolling of a biased die. The likelihood in that case is

$$L(\theta) = 120\,\theta^7(1-\theta)^3.$$

A graph of this function is given in Figure 7, a repeat of Figure 1 but with its maximum clearly marked.

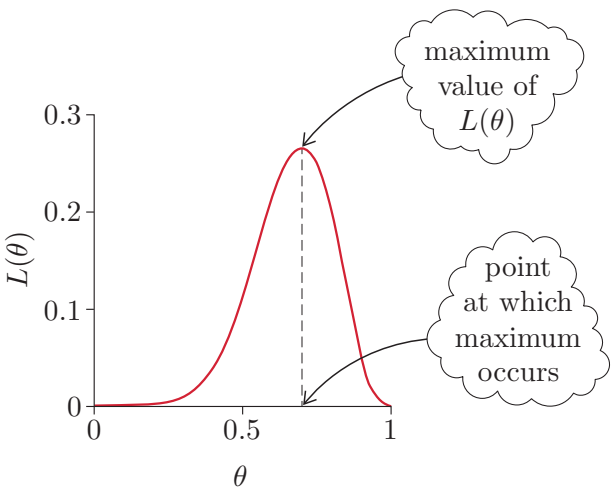

Figure 7 The likelihood $L(\theta)$ for $0 < \theta < 1$ with its maximum marked

Recall that the equation of the slope, or gradient, of a function of one variable is the derivative of that function. For example, the gradient of the likelihood $L(\theta)$, which is a function of the variable θ, is its derivative $L'(\theta)$.

Now, as you can see from Figure 7, the maximum of the function $L(\theta)$ occurs at a point at which the function is (temporarily) flat, that is, where its slope, or gradient, is zero. At such a point – referred to in general as a *stationary point* – the derivative of the curve is zero, that is, $L'(\theta) = 0$.

Figure 7 also illustrates the following important properties that are often possessed by a likelihood.

Property 1. The likelihood $L(\theta)$ is a smooth function of θ (see below).

Property 2. There is exactly one stationary point.

Property 3. The function is increasing before the stationary point – that is, its gradient, and therefore its derivative, is positive before the stationary point – and is decreasing after the stationary point – that is, has negative gradient, and therefore derivative, after the stationary point.

Property 4. The likelihood attains its maximum at the stationary point, so the stationary point is the maximum likelihood estimate.

When the likelihood has Properties 1, 2 and 3, Property 4 follows automatically. Thus checking Properties 1, 2 and 3 will often identify the MLE. Indeed, when trying to find the MLE of a single parameter, this is

the approach that is normally tried first – and it generally works except with models that are quite complex or are unusual in some way.

Although 'smoothness' of a function can be given various mathematical definitions involving its continuity and the behaviour of its derivatives, we need not do so here. In this module, you may assume that all the likelihoods you encounter are smooth, and hence that Property 1 holds.

Property 2 can be more usefully expressed in terms of the derivative of the likelihood function. It is equivalent to:

Property 2*. The equation $L'(\theta) = 0$ has a single solution, say $\theta = \theta^*$.

We can therefore, for example, locate the value of θ^* for the likelihood shown in Figure 7 by solving the equation $L'(\theta) = 0$ when $L(\theta) = 120\,\theta^7(1-\theta)^3$. Using Equation (9) to differentiate a product and Equation (8) to differentiate the second term in the product, we have

$$\begin{aligned} L'(\theta) &= 120\left\{7\theta^6 \times (1-\theta)^3 + \theta^7 \times (-1) \times 3(1-\theta)^2\right\} \\ &= 120\left\{7\theta^6(1-\theta)^3 - 3\theta^7(1-\theta)^2\right\} \\ &= 120\,\theta^6(1-\theta)^2\left\{7(1-\theta) - 3\theta\right\} \\ &= 120\,\theta^6(1-\theta)^2\,(7-10\theta)\,. \end{aligned}$$

Now, since $0 < \theta < 1$, the term $120\,\theta^6(1-\theta)^2$ is positive; call it $K(\theta)$, say. So when we set $L'(\theta) = 0$, we have

$$K(\theta)(7-10\theta) = 0.$$

Thus the value θ^* must satisfy

$$7 - 10\theta^* = 0.$$

This is easily solved to yield

$$\theta^* = \frac{7}{10} = 0.7.$$

What of Property 3? Well, this can be checked, more or less formally, in any of three ways:

- look at the graph of $L(\theta)$
- explicitly check that $L'(\theta) > 0$ for $\theta < \theta^*$ and $L'(\theta) < 0$ for $\theta > \theta^*$
- check an equivalent formulation of Property 3 in terms of the sign of the *second* derivative of $L(\theta)$ (that is, the derivative of the derivative $L'(\theta)$) at θ^*.

You will not be asked to make any of these checks in this module. Indeed, the only thing you need check is that there is exactly one solution of the equation $L'(\theta) = 0$ for 'allowed' values of θ. If so, you can take it that this solution, θ^*, is also the MLE $\widehat{\theta}$. So, for example,

$$\widehat{\theta} = \theta^* = 0.7$$

is the MLE of θ in the example of the rolling of a biased die. (This result was mentioned in Example 8.)

Things that can go wrong with this approach outside this module include that a single stationary point might be a minimum or another kind of stationary point, or that the likelihood has multiple maxima.

The key, therefore, to using differentiation to obtain exact values for MLEs is Property 2*: differentiate $L(\theta)$, and find the solution of $L'(\theta) = 0$. The following box gives the full set of steps to follow when looking for an MLE of a single parameter θ in M248.

Finding the MLE of θ

Step 1. Form the likelihood $L(\theta)$ as a product of simpler terms, as in Equation (7).

Step 2. Differentiate $L(\theta)$ to obtain $L'(\theta)$.

Step 3. Solve the equation $L'(\theta) = 0$. If there is exactly one solution, then set the MLE $\widehat{\theta}$ equal to that solution.

The above procedure is used for both discrete probability models and continuous probability models. Here are some examples and activities.

Example 16 *Estimating the parameter of a geometric distribution again*

In Example 9, a very small artificial dataset of three observations, $x_1 = 3$, $x_2 = 4$ and $x_3 = 8$, was considered. Assuming that these are independent observations from a geometric distribution with unknown parameter θ, $0 < \theta < 1$, the likelihood for θ was shown to be

$$L(\theta) = (1-\theta)^{12}\theta^3.$$

Step 1 of finding the MLE is therefore (already) completed.

Step 2 asks us to obtain $L'(\theta)$. Using Equation (9) to differentiate a product and Equation (8) to differentiate the first term in the product, this is

$$\begin{aligned} L'(\theta) &= (-1) \times 12(1-\theta)^{11} \times \theta^3 + (1-\theta)^{12} \times 3\theta^2 \\ &= -12(1-\theta)^{11}\theta^3 + 3(1-\theta)^{12}\theta^2 \\ &= 3(1-\theta)^{11}\theta^2\{-4\theta + (1-\theta)\} \\ &= 3(1-\theta)^{11}\theta^2(1-5\theta). \end{aligned}$$

Step 3 asks us to solve $L'(\theta) = 0$. Since for $0 < \theta < 1$ we have $3(1-\theta)^{11}\theta^2 > 0$, this reduces to solving the linear equation

$$1 - 5\theta = 0.$$

This obviously has a single solution, namely the MLE

$$\widehat{\theta} = \frac{1}{5} = 0.2.$$

Example 17 *Sparrow nests*

For each of 40 plots of land, the number of sparrow nests in that plot was recorded. The data are given in the following table.

Table 8 Numbers of sparrow nests in each of 40 plots

Number of nests	0	1	2	3	> 3
Observed frequency	9	22	7	2	0

According to a recent British Trust for Ornithology report, the sparrow population in the UK has declined by nearly half since the 1970s

Assume that the number of nests in a plot follows a Poisson distribution, and let θ denote the mean of the distribution; note that $\theta > 0$. We want to obtain the MLE of θ. First, we need the likelihood. The Poisson p.m.f. is

$$p(x;\theta) = \frac{e^{-\theta}\theta^x}{x!}.$$

Thus the likelihood is

$$\begin{aligned} L(\theta) &= p(0;\theta)^9 \times p(1;\theta)^{22} \times p(2;\theta)^7 \times p(3;\theta)^2 \\ &= \underbrace{e^{-\theta} \times \cdots \times e^{-\theta}}_{9 \text{ times}} \times \underbrace{\frac{e^{-\theta}\theta}{1!} \times \cdots \times \frac{e^{-\theta}\theta}{1!}}_{22 \text{ times}} \\ &\quad \times \underbrace{\frac{e^{-\theta}\theta^2}{2!} \times \cdots \times \frac{e^{-\theta}\theta^2}{2!}}_{7 \text{ times}} \times \frac{e^{-\theta}\theta^3}{3!} \times \frac{e^{-\theta}\theta^3}{3!} \\ &= \left(e^{-\theta}\right)^9 \left(\frac{e^{-\theta}\theta}{1!}\right)^{22} \left(\frac{e^{-\theta}\theta^2}{2!}\right)^7 \left(\frac{e^{-\theta}\theta^3}{3!}\right)^2 \\ &= \frac{e^{-9\theta-22\theta-7\theta-2\theta}\theta^{22+14+6}}{2^7 6^2} \\ &= \frac{e^{-40\theta}\theta^{42}}{4608}. \end{aligned}$$

Recall that $e^a e^b \cdots e^k = e^{a+b+\cdots+k}$ and $(e^x)^k = e^{kx}$.

Using Equation (9) to differentiate a product, we have

$$\begin{aligned} L'(\theta) &= \frac{1}{4608}\left(-40e^{-40\theta} \times \theta^{42} + e^{-40\theta} \times 42\theta^{41}\right) \\ &= \frac{1}{2304} e^{-40\theta}\theta^{41}\left(-20\theta + 21\right). \end{aligned}$$

All but the linear term in brackets is irrelevant to solving $L'(\theta) = 0$ because for $\theta > 0$, $e^{-40\theta}\theta^{41}/2304 > 0$. The linear term has a single value of θ at which it is zero, so that value is the MLE $\widehat{\theta}$: $\widehat{\theta}$ satisfies

$e^x > 0$ for any value of x.

$$-20\widehat{\theta} + 21 = 0,$$

so

$$\widehat{\theta} = \frac{21}{20} = 1.05.$$

Activity 16 *Flood frequency*

Example 10 concerned the frequency of floods in a town. It was assumed that the time between floods is an observation from an exponential distribution with parameter $\theta > 0$. The data took values 5, 8 and 7 years.

It was shown that the likelihood for θ is

$$L(\theta) = \theta^3 e^{-20\theta}.$$

(a) Determine $L'(\theta)$.

(b) Hence show that the MLE of θ takes the value 0.15.

Activity 17 *Industrial component inspections*

Inspecting non-industrial components coming off a production line

Table 9 gives the frequencies of counts of the numbers of inspections of batches of industrial components which find no problem up to and including an inspection which finds one or more problems. Notice that counts (such as 6 or 15 inspections) which have no occurrences in the dataset are omitted from the table. In this dataset, $n = 28$.

Table 9 Counts of industrial inspections up to and including one that finds a problem

Count	1	2	3	4	5	7	9	11	13	14	17	18	26	29
Frequency	6	4	3	3	2	1	1	1	1	2	1	1	1	1

(Source: Bracquemond, C., Crétois, E. and Gaudoin, O., 'A comparative study of goodness-of-fit for the geometric distribution and application to discrete time reliability', Undated technical report)

A good model for these data would appear to be a geometric distribution with parameter θ; here, $0 < \theta < 1$.

(a) Show that the likelihood for these data is

$$L(\theta) = (1-\theta)^{175}\theta^{28}.$$

(b) Determine $L'(\theta)$.

(c) Hence find the MLE of θ.

Exercises on Section 3

Exercise 3 *Practice with differentiation*

Find the derivatives of the following functions.

(a) $3 + \dfrac{4}{x} - \dfrac{11}{x^5}$

(b) $\sqrt{1+x^3}$

(c) $\dfrac{(1+x)^{10}}{\sqrt{x}}$

(d) $x^2 e^{-x}$

Exercise 4 *Clay pigeon shooting*

In Exercise 2, the numbers of shots that a man took to hit four clay targets were given (3, 1, 2 and 2) and modelled by a geometric distribution

with parameter θ, $0 < \theta < 1$. The likelihood for θ was shown in Exercise 2(a) to be

$$L(\theta) = \theta^4(1-\theta)^4.$$

(a) Determine $L'(\theta)$.

(b) Hence find the MLE of θ.

Exercise 5 *Tyre machine failure times*

Table 10 gives the times from repair until failure (in hours) of a machine applying coatings to tyres in a factory in Iraq. Here, $n = 22$.

Table 10 Times until failure of machine (in hours)

3.5	6.5	10.5	23.25	23.5	43.5	69	70
75.5	83.25	95.5	109.5	111.25	144	164	167.25
253	383.75	417.75	428.25	453	1215		

(Source: Al-Jammal, Z.Y. (2008) 'Exponentiated exponential distribution as a failure time distribution', *Iraqi Journal of Statistical Science*, vol. 14, pp. 63–75)

A good model for these data would appear to be an exponential distribution with parameter $\theta > 0$.

(a) Show that the likelihood for these data is

$$L(\theta) = \theta^{22}e^{-4350.75\,\theta}.$$

(b) Determine $L'(\theta)$.

(c) Hence find the MLE of θ.

4 Maximum likelihood estimators and their properties

In this section, we use differentiation to find maximum likelihood *estimators* of model parameters (not just specific maximum likelihood *estimates*). That is, we use the maximum likelihood approach to derive estimating formulas for parameters – maximum likelihood estimators – rather than just estimates for specific datasets. After doing so in Subsection 4.1, in Subsection 4.2 we briefly outline some attractive properties of maximum likelihood estimators.

The abbreviation MLE is used to denote both maximum likelihood estimates and maximum likelihood estimators.

4.1 Maximum likelihood estimators

In two examples in Subsection 3.2, we used sample data to determine the maximum likelihood estimate of the parameter of a geometric distribution. The first example (Example 16) was an artificial one with $n = 3$ and data values 3, 4 and 8. For this dataset, the sample mean is

MLE Pyrotechnics of Daventry: providing the fireworks in M248?

$\overline{x} = (3 + 4 + 8)/3 = 15/3 = 5$ and, from Example 16, the MLE of θ is $\widehat{\theta} = 0.2 = 1/5$. So, in that case, the MLE is the reciprocal of the sample mean. The second example (Activity 17) concerned inspections of industrial components; the data were given in Table 9. For this dataset, the sample mean is

$$\overline{x} = \frac{6 \times 1 + 4 \times 2 + 3 \times 3 + \cdots + 1 \times 29}{28} = \frac{203}{28};$$

also, from Activity 17, the MLE of θ is $\widehat{\theta} = 28/203$ which, again, is the reciprocal of the sample mean.

Obviously, it would be helpful to know if the MLE of the parameter of a geometric distribution is *always* equal to the reciprocal of the sample mean. If it is, then for geometric distributions, we would not need to use calculus to find the MLE, but could simply set the MLE equal to the reciprocal of the sample mean. In Example 18, we will show that this is the case. Indeed, for most standard distributions there are known formulas for the maximum likelihood estimators of the parameters of the distribution. Let's investigate what these might be.

To recall the definition of a likelihood, suppose that $X_1, X_2, \ldots, X_n$ is a random sample that takes values $x_1, x_2, \ldots, x_n$. If these values are from a discrete distribution with probability mass function $p(x; \theta)$, then the likelihood of θ is

$$L(\theta) = p(x_1; \theta) \times p(x_2; \theta) \times \cdots \times p(x_n; \theta).$$

Similarly, if they are from a continuous distribution with probability density function $f(x; \theta)$, then

$$L(\theta) = f(x_1; \theta) \times f(x_2; \theta) \times \cdots \times f(x_n; \theta).$$

Remember that an estimator is a formula, and an estimate is its value.

To obtain the maximum likelihood estimator, we first determine the maximum likelihood estimate. This will be a function of the observations $x_1, x_2, \ldots, x_n$. Then the maximum likelihood estimator is obtained simply by replacing observed quantities by the corresponding random variables: the observation x_1 would be replaced by X_1, x_2 by X_2, and so forth.

The approach is illustrated in the next examples and activities.

Example 18 *Maximum likelihood estimator for a geometric distribution*

Suppose that $x_1, x_2, \ldots, x_n$ make up a random sample of observations from a geometric distribution with parameter θ, $0 < \theta < 1$. The likelihood is

$$\begin{aligned} L(\theta) &= p(x_1; \theta) \times p(x_2; \theta) \times \cdots \times p(x_n; \theta) \\ &= (1-\theta)^{x_1 - 1}\theta \times (1-\theta)^{x_2 - 1}\theta \times (1-\theta)^{x_3 - 1}\theta \times \cdots \times (1-\theta)^{x_n - 1}\theta \\ &= (1-\theta)^{\sum_{i=1}^{n} x_i - n}\theta^n. \end{aligned}$$

Now, $\sum_{i=1}^{n} x_i / n = \overline{x}$, so $\sum_{i=1}^{n} x_i = n\overline{x}$. It follows that the likelihood can be written in a slightly simpler form based on the sample mean:

$$L(\theta) = (1-\theta)^{n\overline{x} - n}\theta^n.$$

To obtain $L'(\theta)$, we use Equation (9) to differentiate a product and Equation (8) to differentiate the first term in the product:

$$\begin{aligned} L'(\theta) &= (-1) \times (n\overline{x} - n)(1-\theta)^{n\overline{x}-n-1} \times \theta^n + (1-\theta)^{n\overline{x}-n} \times n\theta^{n-1} \\ &= n(1-\theta)^{n\overline{x}-n-1}\theta^{n-1}\{-(\overline{x}-1)\theta + (1-\theta)\} \\ &= n(1-\theta)^{n\overline{x}-n-1}\theta^{n-1}(1-\overline{x}\theta). \end{aligned}$$

Now, since $0 < \theta < 1$, the multiplier $n(1-\theta)^{n\overline{x}-n-1}\theta^{n-1} > 0$, so the only solution of $L'(\theta) = 0$ is when

$$1 - \overline{x}\theta = 0.$$

The maximum likelihood *estimate* for this set of data is therefore

$$\widehat{\theta} = \frac{1}{\overline{x}}.$$

So, replacing the observed sample mean by its random variable version, the maximum likelihood *estimator* of the parameter θ of a geometric distribution for any set of data is therefore given by

$$\widehat{\theta} = \frac{1}{\overline{X}}.$$

The MLE of the geometric parameter is indeed always equal to the reciprocal of the sample mean.

Example 19 *Maximum likelihood estimator for a Poisson distribution*

Suppose that $x_1, x_2, \ldots, x_n$ is a random sample of observations from a Poisson distribution with parameter θ, $\theta > 0$. To obtain the maximum likelihood estimate of θ, we first form the likelihood:

$$\begin{aligned} L(\theta) &= p(x_1;\theta) \times p(x_2;\theta) \times \cdots \times p(x_n;\theta) \\ &= \frac{e^{-\theta}\theta^{x_1}}{x_1!} \times \frac{e^{-\theta}\theta^{x_2}}{x_2!} \times \cdots \times \frac{e^{-\theta}\theta^{x_n}}{x_n!} \\ &= \frac{e^{-n\theta} \times \theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}. \end{aligned}$$

Notice that $1/\prod_{i=1}^n x_i!$ is positive and does not depend on θ, so can be written as a positive constant, C, say. Also, using the fact that $\sum_{i=1}^n x_i = n\overline{x}$, the likelihood can then be written more simply as

$$L(\theta) = Ce^{-n\theta}\theta^{n\overline{x}}.$$

It follows that, differentiating the product (using Equation (9)),

$$\begin{aligned} L'(\theta) &= C\big(-ne^{-n\theta} \times \theta^{n\overline{x}} + e^{-n\theta} \times n\overline{x}\theta^{n\overline{x}-1}\big) \\ &= Cne^{-n\theta}\theta^{n\overline{x}-1}(-\theta + \overline{x}). \end{aligned}$$

Since θ is positive, so is $Cne^{-n\theta}\theta^{n\overline{x}-1}$, and the only solution of $L'(\theta) = 0$ is when

$$-\theta + \overline{x} = 0,$$

namely

$$\widehat{\theta} = \overline{x}.$$

Again, replacing the observed sample mean by its random variable version, the maximum likelihood estimator of the parameter θ of a Poisson

distribution for any set of data is therefore given by

$$\widehat{\theta} = \overline{X}.$$

That is, the MLE of the Poisson parameter is always equal to the sample mean.

In Example 17, we saw that the maximum likelihood estimate of θ for the data on sparrow nests given in Table 8, assuming a Poisson distribution, is $\widehat{\theta} = 21/20 = 1.05$. Instead of deriving this MLE from scratch, had we known what we now know, we could have declared the MLE to be the sample mean and just calculated the latter. From Table 8,

$$\overline{x} = \frac{9 \times 0 + 22 \times 1 + 7 \times 2 + 2 \times 3}{40} = \frac{42}{40} = \frac{21}{20} = 1.05.$$

You can work out the general form of the MLE for binomial and exponential parameters for yourself in Activities 18 and 19, respectively.

Activity 18 *Maximum likelihood estimator for a binomial distribution*

Suppose that n independent trials are conducted and the number of successes, X, is a random variable that follows the binomial distribution $B(n, \theta)$, $0 < \theta < 1$.

(a) What is the likelihood for θ based on the single observation $X = x$?

(b) Determine the maximum likelihood estimate of θ when $X = x$.

(c) Hence write down the maximum likelihood estimator of θ.

Activity 19 *Maximum likelihood estimator for an exponential distribution*

Suppose that $x_1, x_2, \ldots, x_n$ is a random sample of observations from an exponential distribution with parameter θ, $\theta > 0$, these being the observed values of independent random variables $X_1, X_2, \ldots, X_n$ from this distribution.

(a) What is the likelihood for θ?

(b) Determine the maximum likelihood estimate of θ.

(c) Hence write down the maximum likelihood estimator of θ. How does the maximum likelihood estimator depend on the sample mean?

So we have found, for a number of standard distributions, formulas for the maximum likelihood estimators of the parameters of the distribution. When sample data come from such a distribution, applying these formulas is the quickest and easiest way of obtaining a maximum likelihood estimate – just feed sample values into the formula for the estimator to obtain the estimate.

Table 11 contains a list of standard results for maximum likelihood estimators for the parameters – in their more usual notation, instead of using θ everywhere – of some of the more well-known probability models. In all but one case, these estimators assume a random sample $X_1, X_2, \ldots, X_n$ with sample mean $\overline{X}$. The exception is the binomial distribution, whose estimator is based on a single observation X. The latter could have been made to match the others by thinking of X as being derived (as the total number of successes) from an underlying random sample of n observations from a Bernoulli distribution each with parameter p (n independent Bernoulli trials). For better or worse, Table 11 follows the most standard way of presenting these results in statistics textbooks.

The table also states whether or not the estimator is unbiased, that is, whether or not its average value is equal to the parameter it is estimating. You need not worry about proving all the results about bias; they are included for information only.

Table 11 Maximum likelihood estimators (MLEs) for some standard probability models

Probability distribution	Estimator	Properties
Binomial, $B(n,p)$	$\widehat{p} = X/n$	$E(\widehat{p}) = p$
Geometric, $G(p)$	$\widehat{p} = 1/\overline{X}$	$\widehat{p}$ is biased
Poisson(λ)	$\widehat{\lambda} = \overline{X}$	$E(\widehat{\lambda}) = \lambda$
Exponential, $M(\lambda)$	$\widehat{\lambda} = 1/\overline{X}$	$\widehat{\lambda}$ is biased
Normal, $N(\mu, \sigma^2)$	$\widehat{\mu} = \overline{X}$	$E(\widehat{\mu}) = \mu$
	$\widehat{\sigma}^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2/n$	$\widehat{\sigma}^2$ is biased

The first four MLEs listed in Table 11 were derived in Examples 18 and 19 and Activities 18 and 19. The MLEs of the parameters of the normal distribution will not be derived here, partly because we have not been dealing with maximum likelihood estimation of two parameters at once in this unit. They are, however, important to recognise; for the normal distribution

- the MLE of μ is the sample mean, $\overline{X}$
- the MLE of σ^2 is the estimator

$$\widehat{\sigma}^2 = W = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

(and not the unbiased estimator S^2, which has divisor $n-1$ in place of n).

And this is the DJ MLE

You should be reassured to observe that in almost all of the estimation problems connected with standard models summarised above, the method of maximum likelihood has led back to estimators that are identical to the estimators that we might have come up with on intuitive grounds. In fact, the estimators X/n for the binomial parameter and $\overline{X}$ for both Poisson and normal means were all considered in Section 1 of this unit, before we

embarked on our odyssey into maximum likelihood estimation. Also, the estimators $1/\overline{X}$ were previously suggested in Subsection 1.2 of Unit 4 for the geometric parameter and in Subsection 2.2 of Unit 5 for the exponential parameter.

The following activity requires you to use results from Table 11 to obtain maximum likelihood estimates.

Activity 20 *Maximum likelihood estimates for data from standard probability models*

Given each of the following samples, use Table 11 to determine the maximum likelihood estimate of the unknown parameter.

(a) Observations 5, 3, 19, 9, 4, 7, 8, 3, 15, 5, 7, 4 from the geometric distribution, $G(p)$.

(b) Observations given in Table 1, and considered in Example 1, on counts of the leech *Helobdella* in 103 water samples, assumed to come from the Poisson distribution, Poisson(λ). Hint: no new calculations are needed, so you can reuse a result that was obtained earlier in the unit.

(c) Observations 0.131, 2.58, 0.04, 4.64, 1.70, 0.40, 4.28, 0.19 from the exponential distribution, $M(\lambda)$.

Activity 21 *Maximum likelihood estimates for data from the normal distribution*

(a) In Example 21 of Unit 6, a small sample of $n = 9$ Byzantine coins whose silver content had been measured was considered. (These were the coins from the first of four coinages.) On consideration of a normal probability plot, it was thought plausible that the silver contents could be modelled by a normal, $N(\mu, \sigma^2)$, distribution. It turns out that the sample mean of these data is $\overline{x} \simeq 6.7444\,\%$ and the sample variance is $s^2 \simeq 0.2953\,\%^2$.

(i) What is the maximum likelihood estimate of the parameter μ?

(ii) What is the maximum likelihood estimate of the parameter σ^2? Hint: you will have to work this out from the values of n and s^2.

(b) In Example 3, we revisited data on the chest circumferences of $n = 5732$ nineteenth-century Scottish soldiers (in inches). A normal distribution, $N(\mu, \sigma^2)$, was deemed to be an appropriate model for these data. The sample mean is $\overline{x} \simeq 39.8489$ inches and the sample variance is $s^2 \simeq 4.2989$ inches^2.

(i) What is the maximum likelihood estimate of the parameter μ?

(ii) What is the maximum likelihood estimate of the parameter σ^2?

(c) Comment on the similarity or otherwise between the values of s^2 and the MLE of σ^2 for these two datasets.

4.2 Properties of maximum likelihood estimators

In Subsection 1.2, the question was asked: 'What makes a good estimator?' It was suggested that an estimator is useful if it has low bias, indeed preferably no bias at all – that is, it is unbiased – and has a small variance. For most estimation methods, it is impossible to make general statements about the properties of the estimators they yield; the properties will vary with the underlying probability model. If the sample size is small, the same is true of maximum likelihood estimators. For large sample sizes, however, maximum likelihood estimators possess certain good properties. Results are said to hold **asymptotically** if they are approximately true provided that the sample size is large enough.

The Central Limit Theorem of Unit 6 is an example of an asymptotic result.

The statistical theory behind the results in the following box is difficult, mathematically, and details will not be given here. You should accept these claims on trust.

Properties of maximum likelihood estimators

- Maximum likelihood estimators are sometimes unbiased and typically have small bias. Also, they are **asymptotically unbiased**; that is,

  $$E(\widehat{\theta}) \to \theta \text{ as } n \to \infty,$$

  where n is the sample size. It follows that maximum likelihood estimators are approximately unbiased for large sample sizes.

As before, the symbol '$\to$' is read as 'tends to'.

- In addition, for maximum likelihood estimators, the variance $V(\widehat{\theta})$ tends to 0 with increasing sample size:

  $$V(\widehat{\theta}) \to 0 \text{ as } n \to \infty.$$

  Moreover, for large n, no unbiased estimator of θ has a smaller variance than the maximum likelihood estimator.

So maximum likelihood estimators possess the sorts of useful properties identified in Subsection 1.2. For instance, if they are not (exactly) unbiased for θ, then for large samples they are at least approximately unbiased for θ, and their variance becomes small.

Example 20 *Asymptotic unbiasedness of the MLE of a normal variance*

To illustrate how an estimator's bias can decline with increasing sample size, consider the MLE of the variance (σ^2) of a normal distribution. From Table 11, the MLE is $\widehat{\sigma}^2 = W = \sum(X_i - \overline{X})^2/n$. In Subsection 1.2 of this unit (Activity 5) you showed that

$$E(\widehat{\sigma}^2) = \frac{n-1}{n}\,\sigma^2 = \sigma^2 - \frac{1}{n}\,\sigma^2.$$

Hence the MLE of σ^2 is biased, with a bias of $-\sigma^2/n$. Now, as n increases, $1/n$ decreases, so the bias also gets smaller. Indeed, $1/n \to 0$ as $n \to \infty$, so

$$-\frac{1}{n}\sigma^2 \to 0 \text{ as } n \to \infty,$$

and hence

$$E(\widehat{\sigma}^2) \to \sigma^2 \text{ as } n \to \infty.$$

This is consistent with the observation that $\widehat{\sigma}^2 = W$ differs little from the unbiased estimator S^2 in an example with large n in Activity 21(b).

Thus the MLE of σ^2 is asymptotically unbiased, even though it has some (small) bias for finite sample sizes.

It is also possible to state useful conclusions not just about the mean and variance of maximum likelihood estimators, but about their sampling distribution as well. Most importantly, maximum likelihood estimators are asymptotically normally distributed, provided that certain mild conditions apply. (The conditions are satisfied by most probability models.) However, this kind of result requires a certain amount of supporting theory before it can be confidently applied, and it will not be pursued further in this module.

The Central Limit Theorem is very relevant!

In other texts or modules, you will often see maximum likelihood estimation approached through $\ell(\theta) = \log\{L(\theta)\}$, the so-called log-likelihood. This is very much a valid approach that can make the derivation of MLEs somewhat simpler – provided that you are comfortable with the rules for manipulating and differentiating logarithms.

The log-likelihood can also avoid computational problems with the likelihood, associated with values of the likelihood sometimes being extremely small.

It has been shown that the method of maximum likelihood can often lead to estimators that are identical to the common sense estimators that we might guess without any supporting theory. A benefit of deriving an estimator by maximum likelihood is that it is then known to possess the above desirable properties.

Exercises on Section 4

Exercise 6 *Maximum likelihood estimator for a Rayleigh distribution*

In Activity 10, the Rayleigh distribution was introduced. Its p.d.f. is

$$f(x;\theta) = \frac{x}{\theta^2}\, e^{-x^2/2\theta^2}, \quad x > 0,$$

with $\theta > 0$. Let us suppose that we have available a random sample $x_1, x_2, \ldots, x_n$, these being the observed values of independent random variables $X_1, X_2, \ldots, X_n$ from this distribution. Set

$$C = \prod_{i=1}^{n} x_i > 0, \quad m_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 \quad \text{and} \quad M_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Note that $m_2 \neq \overline{x}^2, M_2 \neq \overline{X}^2$.

(a) Show that the likelihood for θ can be written

$$L(\theta) = C\theta^{-2n} e^{-nm_2\theta^{-2}/2}.$$

(b) Determine the maximum likelihood estimate of θ.

(c) Hence write down the maximum likelihood estimator of θ.

Exercise 7 *Douglas firs*

The ecologist E.C. Pielou was interested in the pattern of healthy and diseased trees in a plantation of Douglas firs. (The disease that was the subject of her research was '*Armillaria* root rot'.) Several narrow lines of trees through the plantation (called 'transects') were examined. After each diseased tree, X, the number of trees that had to be examined in order to find a healthy tree was counted. There were 109 such counts. Their frequency distribution is given in Table 12.

Table 12 Numbers of trees examined to find a healthy tree

Number of trees, X	1	2	3	4	5	6
Frequency	71	28	5	2	2	1

(Source: Pielou, E.C. (1963) 'Runs of healthy and diseased trees in transects through an infected forest', *Biometrics*, vol. 19, no. 4, pp. 603–14)

Douglas firs in Canada

(a) Assuming that X has a geometric distribution with parameter p, use information given in Table 11 to determine the maximum likelihood estimate of p.

(b) In fact, information from a total of 166 trees is summarised in Table 12. Of these trees, 109 were healthy and 57 were unhealthy. Let Y be a random variable representing the number of healthy trees in a collection of 166 trees. Suppose that $Y \sim B(166, p)$. Use information given in Table 11 to write down the maximum likelihood estimate of p.

(c) Say whether each of the estimators in parts (a) and (b) is biased. Does this result surprise you?

Summary

There are various ways of obtaining estimating formulas – that is, *estimators* – of unknown model parameters; when an estimating formula is applied in a data context, the resulting number provides an *estimate* of the unknown parameter. The quality of an estimate can be assessed from the properties of the sampling distribution of the corresponding estimator. Not all estimating procedures are applicable in all data contexts, and not all estimating procedures are guaranteed always to give sensible estimates. Often, though, reasonable estimation methods give estimates that are either identical or very similar to one another. Qualities in an estimator that are desirable include unbiasedness and small variance.

One particular estimation method has been discussed in detail in this unit – the method of maximum likelihood. Maximum likelihood is one of the most useful estimation methods available because of its versatility and because it leads to estimators that have good properties. In particular, maximum likelihood estimators are both asymptotically unbiased and have variance tending to zero. The maximum likelihood estimate is the value of the parameter that is most likely to give rise to the observed data as measured by the likelihood function, or just likelihood. In many situations, the MLE can be obtained by taking the derivative of the likelihood and equating the derivative to zero. For most standard distributions, there are known formulas for the maximum likelihood estimators of the distribution's parameters.

Learning outcomes

After you have worked through this unit, you should be able to:

- realise that a parameter can have a variety of plausible estimators
- compare estimators in simple contexts by examining their means and variances
- understand the notion of unbiasedness and that it is a desirable quality in an estimator
- appreciate that the sample variance is an unbiased estimator of the variance of any (not necessarily normal) distribution
- acknowledge that if two estimators are both unbiased, then the one with the smaller variance is usually preferred, but also that estimators with both small bias and small variance can be useful
- appreciate that the method of maximum likelihood is an important way of estimating a parameter
- construct the likelihood $L(\theta)$ associated with random samples from simple models
- determine maximum likelihood estimates and estimators in well-behaved one-parameter problems through differentiation; this involves obtaining $L'(\theta)$ and choosing the MLE $\widehat{\theta}$ to solve the equation $L'(\theta) = 0$
- use standard results to obtain maximum likelihood estimates for common probability models
- recognise that desirable qualities in estimators such as asymptotic unbiasedness and variance tending to zero are possessed by maximum likelihood estimators.

Solutions to activities

Solution to Activity 1

When the population is Poisson(λ), the variance σ^2 is equal to the mean μ, and both are equal to λ. So if a random variable $\overline{W}$ was based on a dataset of size n from the Poisson distribution with parameter λ, then

$$E(\overline{W}) = \lambda, \quad V(\overline{W}) = \frac{\lambda}{n}.$$

Now, the random variable $\overline{X}$ was based on a dataset of size 103, the random variable $\overline{Y}$ on a dataset of size 48. So

$$E(\overline{X}) = E(\overline{Y}) = \lambda,$$

but

$$V(\overline{X}) = \frac{\lambda}{103}, \quad V(\overline{Y}) = \frac{\lambda}{48}.$$

The former is smaller than the latter. This is a particular example of a phenomenon you saw in Unit 6: the larger the sample that is taken, the smaller is the variance of the sample mean.

Solution to Activity 2

For the binomial distribution, $E(X) = np$. Thus

$$E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p.$$

Therefore X/n is an unbiased estimator of p.

Solution to Activity 3

(a) For any probability model, the sample mean is an unbiased estimator of the population mean. So the sample mean is an unbiased estimator of μ.

The variance of the sample mean is σ^2/n (Subsection 6.1 of Unit 6), which is 25/12 in this case.

(b) As n increases, σ^2/n decreases, so the variance of the sample mean would decrease – and the sample mean becomes a better estimator of the population mean – if the sample size were increased.

Solution to Activity 4

(a) $E(\widehat{\mu}_3) = E\{\frac{1}{11}(6X_1 + 3X_2 + 2X_3)\}$
$= \frac{1}{11}E(6X_1 + 3X_2 + 2X_3)$
$= \frac{1}{11}\{E(6X_1) + E(3X_2) + E(2X_3)\}$
$= \frac{1}{11}\{6E(X_1) + 3E(X_2) + 2E(X_3)\}$
$= \frac{1}{11}(6\mu + 3\mu + 2\mu) = \mu.$

Hence $\widehat{\mu}_3$ is an unbiased estimator of μ.

(b) $V(\widehat{\mu}_3) = V\{\frac{1}{11}(6X_1 + 3X_2 + 2X_3)\}$
$= \left(\frac{1}{11}\right)^2 V(6X_1 + 3X_2 + 2X_3)$
$= \left(\frac{1}{11}\right)^2 \{V(6X_1) + V(3X_2) + V(2X_3)\}$
$= \left(\frac{1}{11}\right)^2 \{6^2\,V(X_1) + 3^2\,V(X_2) + 2^2\,V(X_3)\}$
$= \frac{1}{121}(36 \times 1 + 9 \times 4 + 4 \times 9) = \frac{108}{121} \simeq 0.89.$

This is greater than $V(\widehat{\mu}_2) \simeq 0.73$.

(c) The estimator $\widehat{\mu}_2$ is preferred to $\widehat{\mu}_3$ because, while both are unbiased estimators of μ, the variance of $\widehat{\mu}_2$ is smaller than the variance of $\widehat{\mu}_3$.

Solution to Activity 5

(a) $$E(W) = E\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right\} = \frac{1}{n}E\left\{\sum_{i=1}^{n}(X_i - \overline{X})^2\right\}$$
$$= \frac{1}{n} \times (n-1)\sigma^2 \neq \sigma^2,$$

so W is a biased estimator of σ^2.

(b) The bias of W is

$$E(W) - \sigma^2 = \frac{n-1}{n}\,\sigma^2 - \sigma^2 = \frac{n-1-n}{n}\,\sigma^2 = -\frac{1}{n}\,\sigma^2.$$

The estimator W is therefore negatively biased, meaning that on average, W tends to underestimate the value of σ^2.

Solution to Activity 6

(a) The number of mice with adenomas in a sample of size 54 has a binomial distribution $B(54, \theta)$. Given that six mice in the sample had adenomas, the likelihood of θ is

$$L(\theta) = p(6;\theta) = \binom{54}{6}\,\theta^6(1-\theta)^{48} = 25\,827\,165\;\theta^6(1-\theta)^{48}.$$

(b) $L(0.11) = 25\,827\,165\;(0.11)^6(0.89)^{48} \simeq 0.1703,$
$L(0.12) = 25\,827\,165\;(0.12)^6(0.88)^{48} \simeq 0.1669.$

These give the following table.

Table 13

θ	0.09	0.10	0.11	0.12	0.13
$L(\theta)$	0.1484	0.1643	0.1703	0.1669	0.1558

(c) A graph of $L(\theta)$ is shown in Figure 8.

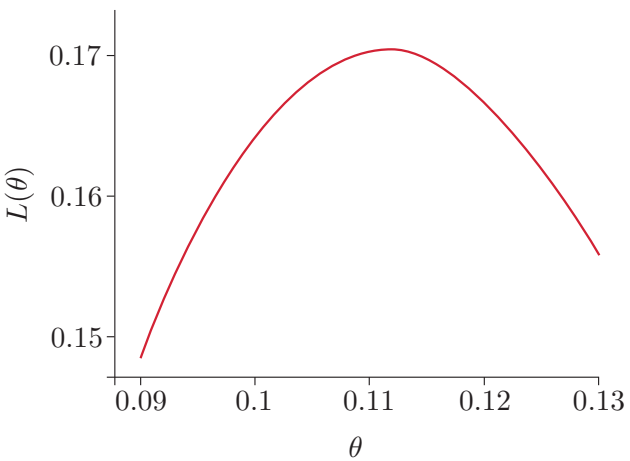

Figure 8 A graph of $L(\theta)$ for $0.09 \leq \theta \leq 0.13$

(d) From the position of the peak of the curve, $\widehat{\theta}$ is a little greater than 0.11, but much smaller than 0.12. So $\widehat{\theta} \simeq 0.11$. (The exact value of $\widehat{\theta}$ is, in fact, $\frac{1}{9} \simeq 0.111$.)

Solution to Activity 7

(a) The likelihood for this random sample is given by

$$\begin{aligned} L(\theta) &= p(x_1;\theta) \times p(x_2;\theta) \times p(x_3;\theta) \times p(x_4;\theta) \\ &= (1-\theta)^{1-1}\theta \times (1-\theta)^{2-1}\theta \times (1-\theta)^{1-1}\theta \times (1-\theta)^{3-1}\theta \\ &= (1-\theta)^{0+1+0+2}\theta^4 \\ &= (1-\theta)^3\theta^4, \end{aligned}$$

as required.

(b) From Figure 3, the likelihood appears to be maximised when θ is just below 0.6. That is, $\widehat{\theta} \simeq 0.58$, say. (The exact MLE turns out to be $\frac{4}{7} \simeq 0.571$ in this case.)

Solution to Activity 8

(a) Each of the 187 instances of unvariegated and unfaded (denoted 0) offspring plants contributes $p(0;\theta)$ to the likelihood, resulting in $p(0;\theta)^{187}$. Similarly, each of the 37 instances of variegated and unfaded (v) offspring plants contributes $p(v;\theta)$ to the likelihood, resulting in $p(v;\theta)^{37}$. And so on. The complete likelihood of θ for the sample observed is therefore given by

$$\begin{aligned} L(\theta) &= p(0;\theta)^{187} \times p(v;\theta)^{37} \times p(f;\theta)^{35} \times p(vf;\theta)^{31} \\ &= \left(\frac{9}{16}+\theta\right)^{187}\left(\frac{3}{16}-\theta\right)^{37}\left(\frac{3}{16}-\theta\right)^{35}\left(\frac{1}{16}+\theta\right)^{31} \\ &= \left(\frac{9}{16}+\theta\right)^{187}\left(\frac{3}{16}-\theta\right)^{72}\left(\frac{1}{16}+\theta\right)^{31}. \end{aligned}$$

(b) From Figure 4, $L(\theta)$ is maximised when θ is somewhere above 0.05; whichever value you chose corresponding to this is the approximate MLE of θ. (A more precise estimate turns out to be $\widehat{\theta} \simeq 0.0584$.)

Solution to Activity 9

(a) For the random sample $x_1 = 25$, $x_2 = 31$ from the exponential distribution, the likelihood of θ is

$$\begin{aligned} L(\theta) &= f(x_1;\theta) \times f(x_2;\theta) = \theta e^{-\theta 25} \times \theta e^{-\theta 31} \\ &= \theta^2 e^{-\theta(25+31)} = \theta^2 e^{-56\theta}. \end{aligned}$$

(b) Figure 6 shows that the MLE of θ is around perhaps 0.04. (Exactly, the maximiser turns out to be $\widehat{\theta} = \frac{1}{28} \simeq 0.0357$.)

Solution to Activity 10

(a) The likelihood $L(\theta)$ is obtained as the product of values of the p.d.f.:

$$\begin{aligned} L(\theta) &= f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_6;\theta) \\ &= f(22.2;\theta) \times f(2.8;\theta) \times \cdots \times f(8.3;\theta) \\ &= \frac{22.2}{\theta^2}\, e^{-22.2^2/2\theta^2} \times \frac{2.8}{\theta^2}\, e^{-2.8^2/2\theta^2} \times \frac{4.0}{\theta^2}\, e^{-4.0^2/2\theta^2} \\ &\quad \times \frac{13.9}{\theta^2}\, e^{-13.9^2/2\theta^2} \times \frac{11.7}{\theta^2}\, e^{-11.7^2/2\theta^2} \times \frac{8.3}{\theta^2}\, e^{-8.3^2/2\theta^2} \\ &\simeq \frac{335\,621}{\theta^{12}}\, e^{-457.8/\theta^2}, \end{aligned}$$

as required.

(b) $L(8.5) = \dfrac{335\,621}{(8.5)^{12}}\, e^{-457.8/(8.5)^2} \simeq 4.178 \times 10^{-9}$,

$L(9) = \dfrac{335\,621}{9^{12}}\, e^{-457.8/9^2} \simeq 4.172 \times 10^{-9}$.

These calculations complete the following table.

Table 14

θ	8.25	8.50	8.75	9.00	9.25
$L(\theta) \times 10^9$	4.048	4.178	4.216	4.172	4.059

(c) A graph of $L(\theta) \times 10^9$ is shown in Figure 9.

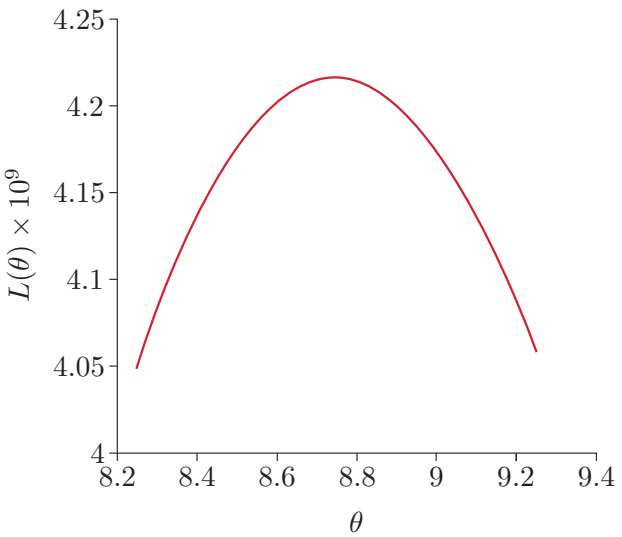

Figure 9 A graph of $L(\theta) \times 10^9$ for $8.25 \leq \theta \leq 9.25$

From the position of the peak of the curve, $\widehat{\theta} \simeq 8.75$. (A more accurate value is $\widehat{\theta} = 8.735$, correct to three decimal places.)

Solution to Activity 11

In each case, let $f(x)$ denote the function to be differentiated. The solutions below, and in other activities in this subsection, give all calculation steps in detail, but you can combine steps to shorten calculations if you are comfortable doing that.

(a) $f'(x) = 2 \times 6x^{2-1} = 12x$.

(b) $f'(x) = 5.1 \times 4x^{5.1-1} = 20.4x^{4.1}$.

(c) Since $f(x) = 5x^{-4}$,

$$f'(x) = -4 \times 5x^{-4-1} = -20x^{-5} = -\frac{20}{x^5}.$$

(d) Since $f(x) = -4x^{3/2}$,

$$f'(x) = \frac{3}{2} \times \left(-4x^{(3/2)-1}\right) = -6x^{1/2} = -6\sqrt{x}.$$

(e) $f'(x) = 27$.

(f) Since $f(x) = -3x^{-1/2}$,

$$f'(x) = -\frac{1}{2} \times \left(-3x^{-(1/2)-1}\right) = \frac{3}{2}\,x^{-3/2} = \frac{3}{2x^{3/2}}.$$

Solution to Activity 12

(a) $$\begin{aligned} f'(x) &= \frac{d}{dx}(4) + \frac{d}{dx}(3x) + \frac{d}{dx}(x^2) + \frac{d}{dx}(-5x^3) + \frac{d}{dx}(2x^7) \\ &= 0 + 3 + 2x^{2-1} + 3 \times (-5x^{3-1}) + 7 \times 2x^{7-1} \\ &= 3 + 2x - 15x^2 + 14x^6. \end{aligned}$$

(b) Since $f(x) = 4 - 3x^{-1/2} - 2x^{-3}$,

$$\begin{aligned} f'(x) &= \frac{d}{dx}(4) - \frac{d}{dx}(3x^{-1/2}) - \frac{d}{dx}(2x^{-3}) \\ &= 0 - \left(-\frac{1}{2}\right) \times 3x^{-(1/2)-1} - (-3) \times 2x^{-3-1} \\ &= \frac{3}{2}\, x^{-3/2} + 6x^{-4} = \frac{3}{2x^{3/2}} + \frac{6}{x^4}. \end{aligned}$$

Solution to Activity 13

(a) $f'(x) = \frac{1}{2} \times 6e^{x/2} = 3e^{x/2}$.

(b) $f'(x) = -3 \times 3e^{-3x} = -9e^{-3x}$.

(c) $f'(x) = -0.1 \times 10e^{-0.1x} = -e^{-0.1x}$.

Solution to Activity 14

(a) Here,

$$h(y) = y^k \quad \text{and} \quad y = g(x) = 1 - x.$$

For these functions,

$$h'(y) = ky^{k-1} \quad \text{and} \quad g'(x) = -1.$$

Therefore

$$f'(x) = -1 \times ky^{k-1} = -1 \times k(1-x)^{k-1} = -k(1-x)^{k-1}.$$

(b) Here,

$$h(y) = 12y^4 \quad \text{and} \quad y = g(x) = 1 + 2x + 2x^{-1/2}.$$

For these functions,

$$h'(y) = 48y^3 \quad \text{and} \quad g'(x) = 2 - x^{-3/2}.$$

Therefore

$$\begin{aligned} f'(x) &= (2 - x^{-3/2}) \times 48y^3 \\ &= (2 - x^{-3/2}) \times 48(1 + 2x + 2x^{-1/2})^3 \\ &= 48\left(2 - \frac{1}{x^{3/2}}\right)\left(1 + 2x + \frac{2}{\sqrt{x}}\right)^3. \end{aligned}$$

Solution to Activity 15

(a) Here,

$$g(x) = x^2 \quad \text{and} \quad h(x) = (1-x)^3.$$

For these functions, $g'(x) = 2x$ and, by the chain rule,

$$h'(x) = -1 \times 3(1-x)^2 = -3(1-x)^2.$$

Therefore

$$\begin{aligned} f'(x) &= 2x \times (1-x)^3 + x^2 \times \{-3(1-x)^2\} \\ &= x(1-x)^2\{2(1-x) - 3x\} = x(1-x)^2(2-5x). \end{aligned}$$

(b) Here,

$$g(x) = x \quad \text{and} \quad h(x) = e^{-x}.$$

For these functions,

$$g'(x) = 1 \quad \text{and} \quad h'(x) = -e^{-x}.$$

Therefore

$$f'(x) = 1 \times e^{-x} + x \times (-e^{-x}) = e^{-x}(1-x).$$

Solution to Activity 16

(a) Using Equation (9) to differentiate a product, we have

$$L'(\theta) = 3\theta^2 \times e^{-20\theta} + \theta^3 \times (-20)e^{-20\theta} = \theta^2 e^{-20\theta}\,(3 - 20\theta)\,.$$

(b) Since $\theta^2 e^{-20\theta} > 0$ for any value of θ, the only solution of $L'(\theta) = 0$ satisfies

$$3 - 20\theta = 0.$$

Therefore the MLE is $\widehat{\theta} = \frac{3}{20} = 0.15$, as required.

Solution to Activity 17

(a) Since the p.m.f. of the geometric distribution is

$$p(x;\theta) = (1-\theta)^{x-1}\theta,$$

the likelihood is

$$\begin{aligned} L(\theta) &= p(1;\theta)^6 \times p(2;\theta)^4 \times p(3;\theta)^3 \times p(4;\theta)^3 \\ &\quad \times \cdots \times p(26;\theta)^1 \times p(29;\theta)^1 \\ &= \underbrace{\theta \times \cdots \times \theta}_{6 \text{ times}} \times \underbrace{(1-\theta)\theta \times \cdots \times (1-\theta)\theta}_{4 \text{ times}} \\ &\quad \times \underbrace{(1-\theta)^2\theta \times \cdots \times (1-\theta)^2\theta}_{3 \text{ times}} \times \underbrace{(1-\theta)^3\theta \times \cdots \times (1-\theta)^3\theta}_{3 \text{ times}} \\ &\quad \times \cdots \times (1-\theta)^{25}\theta \times (1-\theta)^{28}\theta \\ &= (1-\theta)^{4+3\times 2+3\times 3+\cdots+25+28}\,\theta^{6+4+3+3+\cdots+1+1} = (1-\theta)^{175}\theta^{28}, \end{aligned}$$

as required.

(b) Using Equation (9) to differentiate a product and Equation (8) to differentiate the first term in the product, we have

$$\begin{aligned} L'(\theta) &= (-1) \times 175(1-\theta)^{174} \times \theta^{28} + (1-\theta)^{175} \times 28\theta^{27} \\ &= (1-\theta)^{174}\theta^{27}\left\{-175\theta + 28(1-\theta)\right\} \\ &= (1-\theta)^{174}\theta^{27}\left(28 - 203\,\theta\right). \end{aligned}$$

(c) Since $0 < \theta < 1$, we have $(1-\theta)^{174}\theta^{27} > 0$, so the (only) solution of $L'(\theta) = 0$ is the solution of

$$28 - 203\,\theta = 0.$$

Hence

$$\widehat{\theta} = \frac{28}{203} \simeq 0.138.$$

Solution to Activity 18

(a) The likelihood based on a single observation x from a discrete distribution is $L(\theta) = p(x;\theta)$. Using the formula for a binomial probability, we therefore have that

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

(b) To simplify the ensuing formulas, write $C = \binom{n}{x}$ and note that $C > 0$; this is valid because $\binom{n}{x}$ does not depend on θ. This enables us to write

$$L(\theta) = C\theta^x(1-\theta)^{n-x}.$$

Then, differentiating the product (using Equation (9)) and using the chain rule (using Equation (8)) to differentiate the second term in the product,

$$\begin{aligned} L'(\theta) &= C\left\{x\theta^{x-1} \times (1-\theta)^{n-x} + \theta^x \times (-1) \times (n-x)(1-\theta)^{n-x-1}\right\} \\ &= C\theta^{x-1}(1-\theta)^{n-x-1}\{x(1-\theta) - (n-x)\theta\} \\ &= C\theta^{x-1}(1-\theta)^{n-x-1}(x - n\theta). \end{aligned}$$

Since $0 < \theta < 1$, we have $C\theta^{x-1}(1-\theta)^{n-x-1} > 0$, so the only solution of $L'(\theta) = 0$ is the solution of

$$x - n\theta = 0.$$

Therefore the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{x}{n}.$$

(c) Replacing x by X, the maximum likelihood estimator of θ is

$$\widehat{\theta} = \frac{X}{n}.$$

Solution to Activity 19

(a) The likelihood is

$$\begin{aligned} L(\theta) &= f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta) \\ &= \theta e^{-\theta x_1} \times \theta e^{-\theta x_2} \times \cdots \times \theta e^{-\theta x_n} \\ &= \theta^n e^{-\theta(x_1+x_2+\cdots+x_n)} \\ &= \theta^n e^{-\theta \sum_{i=1}^n x_i} = \theta^n e^{-\theta\, n\overline{x}}. \end{aligned}$$

(b) Using Equation (9) to differentiate a product,

$$\begin{aligned} L'(\theta) &= n\theta^{n-1} \times e^{-\theta\, n\overline{x}} + \theta^n \times (-n\overline{x}) e^{-\theta\, n\overline{x}} \\ &= n\theta^{n-1} e^{-\theta\, n\overline{x}}(1 - \theta\overline{x}). \end{aligned}$$

Since θ is positive, so is $n\theta^{n-1}e^{-\theta\, n\overline{x}}$, and the only solution of $L'(\theta) = 0$ is when

$$1 - \theta\overline{x} = 0,$$

that is, the maximum likelihood estimate of θ is

$$\widehat{\theta} = \frac{1}{\overline{x}}.$$

(c) The maximum likelihood estimator of θ for the exponential distribution is

$$\widehat{\theta} = \frac{1}{\overline{X}};$$

it is the reciprocal of the sample mean.

Solution to Activity 20

(a) From Table 11, the maximum likelihood estimate of p is $\widehat{p} = 1/\overline{x}$. The sample mean is

$$\overline{x} = \frac{5 + 3 + 19 + \cdots + 4}{12} = \frac{89}{12} \simeq 7.417.$$

So the MLE of p is $\widehat{p} = \frac{12}{89} \simeq 0.135$.

(b) From Table 11, the maximum likelihood estimate of λ is $\widehat{\lambda} = \overline{x}$. From Example 1, $\overline{x} \simeq 0.816$. So the MLE of λ is $\widehat{\lambda} \simeq 0.816$.

(c) From Table 11, the maximum likelihood estimate of λ is $\widehat{\lambda} = 1/\overline{x}$. The sample mean is

$$\overline{x} = \frac{0.131 + 2.58 + \cdots + 0.19}{8} = \frac{13.961}{8} \simeq 1.745.$$

So the MLE of λ is $\widehat{\lambda} = \frac{8}{13.961} \simeq 0.573$.

Solution to Activity 21

(a) (i) From Table 11, the maximum likelihood estimate of μ is $\widehat{\mu} = \overline{x} \simeq 6.7444\,\%$.

(ii) From Table 11, the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = W = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Recall also that

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Multiplying and dividing the formula for W by $n-1$, it follows that

$$W = \frac{n-1}{n}\,\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n-1}{n}s^2,$$

and the MLE of σ^2 is

$$\widehat{\sigma}^2 = W \simeq \frac{8}{9} \times 0.2953 \simeq 0.2625\,\%^2.$$

(b) (i) The maximum likelihood estimate of μ is $\widehat{\mu} = \overline{x} \simeq 39.8489$ inches.

(ii) As in part (a)(ii), the maximum likelihood estimate of σ^2 is

$$\widehat{\sigma}^2 = W = \frac{n-1}{n}s^2,$$

so

$$\widehat{\sigma}^2 \simeq \frac{5731}{5732} \times 4.2989 \simeq 4.2982 \text{ inches}^2.$$

(c) The values of s^2 and $\widehat{\sigma}^2$ are fairly similar for the coinage data and very similar for the chest measurement data: for the coins, s^2 and $\widehat{\sigma}^2$ differ by 0.0328; for the chest measurements, s^2 and $\widehat{\sigma}^2$ differ by only 0.0007. (The means of both datasets are of broadly comparable size, so it is meaningful to compare the sizes of these differences across datasets.) The degree of similarity between s^2 and $\widehat{\sigma}^2$ is driven by the sample size. For the coin data, n is relatively small and the factor, $(n-1)/n$, by which s^2 is multiplied to obtain $\widehat{\sigma}^2$ is noticeably different from 1 (it is $8/9 \simeq 0.8889$); for the chest measurement data, n is pretty large and the factor by which s^2 is multiplied to obtain $\widehat{\sigma}^2$ is very close to 1 (it is $5731/5732 \simeq 0.9998$).

Solutions to exercises

Solution to Exercise 1

(a) $\overline{X}_1 - \overline{X}_2$ is an unbiased estimator of $\mu_1 - \mu_2$ because

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2,$$

the individual sample means being unbiased estimators of the corresponding population means.

See Subsection 3.3 of Unit 6.

(b) The variance of $\overline{X}_1 - \overline{X}_2$ is

$$\begin{aligned} V(\overline{X}_1 - \overline{X}_2) &= V(\overline{X}_1) + V(\overline{X}_2) \\ &= \frac{9.5^2}{11} + \frac{8.2^2}{10} \simeq 14.93. \end{aligned}$$

See Subsection 3.3 of Unit 6, noting that $\overline{X}_1$ and $\overline{X}_2$ are independent.

(c) If the additional pig were put on the high-protein diet (making twelve pigs on that diet), then the variance of the estimator would be

$$V(\overline{X}_1 - \overline{X}_2) = \frac{9.5^2}{12} + \frac{8.2^2}{10} \simeq 14.24.$$

Alternatively, if the additional pig were put on the low-protein diet (making eleven pigs on that diet), then the variance of the estimator would be

$$V(\overline{X}_1 - \overline{X}_2) = \frac{9.5^2}{11} + \frac{8.2^2}{11} \simeq 14.32.$$

It is good for an estimator to have a small variance – the smaller, the better. So, from this point of view, the extra pig should be put on the high-protein diet.

Solution to Exercise 2

(a) The probability mass function for the geometric distribution with parameter θ is

$$p(x;\theta) = (1-\theta)^{x-1}\theta, \quad x = 1, 2, \ldots.$$

So the likelihood for this particular random sample of size 4 is given by

$$\begin{aligned} L(\theta) &= p(3;\theta) \times p(1;\theta) \times p(2;\theta) \times p(2;\theta) \\ &= (1-\theta)^{3-1}\theta \times (1-\theta)^{1-1}\theta \times (1-\theta)^{2-1}\theta \times (1-\theta)^{2-1}\theta \\ &= (1-\theta)^{2+0+1+1}\theta^4 \\ &= (1-\theta)^4\theta^4. \end{aligned}$$

(b) $L(0.4) = (1-0.4)^4(0.4)^4 = (0.6)^4(0.4)^4 \simeq 0.0033,$
$L(0.5) = (1-0.5)^4(0.5)^4 = (0.5)^8 \simeq 0.0039.$

The completed table for $L(\theta)$ is given below.

Table 15

θ	0.3	0.4	0.5	0.6	0.7
$L(\theta)$	0.0019	0.0033	0.0039	0.0033	0.0019

(c) A graph of $L(\theta)$ is shown in Figure 10.

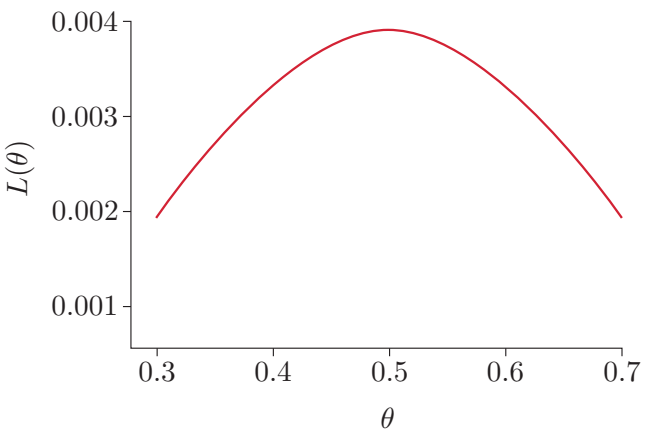

Figure 10 A graph of $L(\theta)$ for $0.3 \leq \theta \leq 0.7$

(d) From the position of the peak of the curve, $\widehat{\theta} \simeq 0.5$. (In fact, 0.5 is the exact value of $\widehat{\theta}$.)

Solution to Exercise 3

(a) Since $f(x) = 3 + 4x^{-1} - 11x^{-5}$,

$$f'(x) = 0 - 4x^{-2} + 55x^{-6} = -\frac{4}{x^2} + \frac{55}{x^6}.$$

(b) Since $f(x) = (1 + x^3)^{1/2}$, this is of the form of Equation (8) if we set

$$h(y) = y^{1/2} \quad \text{and} \quad y = g(x) = 1 + x^3.$$

Now,

$$h'(y) = \frac{1}{2}y^{-1/2} \quad \text{and} \quad g'(x) = 3x^2.$$

It follows that

$$f'(x) = 3x^2 \times \frac{1}{2}y^{-1/2} = 3x^2 \times \frac{1}{2}(1 + x^3)^{-1/2} = \frac{3x^2}{2\sqrt{1 + x^3}}.$$

(c) This function can be written as $g(x)\,h(x)$ where

$$g(x) = x^{-1/2} \quad \text{and} \quad h(x) = (1 + x)^{10}.$$

Now,

$$g'(x) = -\frac{1}{2}x^{-3/2}$$

and, by the chain rule (using Equation (8)),

$$h'(x) = 1 \times 10(1 + x)^9 = 10(1 + x)^9.$$

It follows from Equation (9) that

$$f'(x) = -\frac{1}{2}x^{-3/2} \times (1 + x)^{10} + x^{-1/2} \times 10(1 + x)^9.$$

Noting that

$$x^{-1/2} = \frac{1}{\sqrt{x}} = \frac{x}{x^{3/2}},$$

$f'(x)$ can be simplified to

$$f'(x) = \frac{(1+x)^9}{x^{3/2}}\{-\tfrac{1}{2}(1+x) + 10x\} = \frac{(1+x)^9}{2x^{3/2}}(19x - 1).$$

(d) This function can be written as $g(x)\,h(x)$ where

$$g(x) = x^2 \quad \text{and} \quad h(x) = e^{-x}.$$

Now,

$$g'(x) = 2x \quad \text{and} \quad h'(x) = -e^{-x}.$$

It follows from Equation (9) that

$$f'(x) = 2x \times e^{-x} + x^2 \times (-e^{-x}) = xe^{-x}(2 - x).$$

Solution to Exercise 4

(a) Using Equation (9) to differentiate a product and Equation (8) to differentiate the second term in the product, we have

$$\begin{aligned} L'(\theta) &= 4\theta^3 \times (1-\theta)^4 + \theta^4 \times (-1) \times 4(1-\theta)^3 \\ &= 4\theta^3(1-\theta)^3\{(1-\theta) - \theta\} = 4\theta^3(1-\theta)^3(1-2\theta). \end{aligned}$$

(b) Since $0 < \theta < 1$, we have $4\theta^3(1-\theta)^3 > 0$, so the (only) solution of $L'(\theta) = 0$ is the solution of

$$1 - 2\theta = 0.$$

Hence

$$\widehat{\theta} = \frac{1}{2} = 0.5.$$

Solution to Exercise 5

(a) Since the p.d.f. of the exponential distribution is

$$f(x;\theta) = \theta e^{-\theta x},$$

the likelihood is

$$\begin{aligned} L(\theta) &= \theta e^{-\theta\,3.5} \times \theta e^{-\theta\,6.5} \times \cdots \times \theta e^{-\theta\,1215} \\ &= \theta^{22} e^{-\theta(3.5+6.5+\cdots+1215)} = \theta^{22} e^{-4350.75\,\theta}, \end{aligned}$$

as required.

(b) Using Equation (9) to differentiate a product,

$$\begin{aligned} L'(\theta) &= 22\theta^{21} \times e^{-4350.75\,\theta} + \theta^{22} \times (-4350.75) e^{-4350.75\,\theta} \\ &= \theta^{21} e^{-4350.75\,\theta}(22 - 4350.75\,\theta). \end{aligned}$$

(c) Since $\theta > 0$, we have $\theta^{21} e^{-4350.75\,\theta} > 0$, so the (only) solution of $L'(\theta) = 0$ is the solution of

$$22 - 4350.75\,\theta = 0.$$

Hence

$$\widehat{\theta} = \frac{22}{4350.75} \simeq 0.0051.$$

Solution to Exercise 6

(a) The likelihood is

$$\begin{aligned} L(\theta) &= p(x_1;\theta) \times p(x_2;\theta) \times \cdots \times p(x_n;\theta) \\ &= \frac{x_1}{\theta^2}\, e^{-x_1^2/2\theta^2} \times \frac{x_2}{\theta^2}\, e^{-x_2^2/2\theta^2} \times \cdots \times \frac{x_n}{\theta^2}\, e^{-x_n^2/2\theta^2} \\ &= \frac{x_1 x_2 \cdots x_n}{\theta^{2n}}\, e^{-\sum_{i=1}^n x_i^2/2\theta^2}. \end{aligned}$$

Using the notation C and m_2 defined in the question, and observing that $\sum_{i=1}^n x_i^2 = nm_2$, we have

$$L(\theta) = C\theta^{-2n} e^{-nm_2\theta^{-2}/2},$$

as required.

(b) The final term in the product that makes up $L(\theta)$ can be differentiated using the chain rule (Equation (8)), giving

$$\frac{d}{d\theta} e^{-nm_2\theta^{-2}/2} = nm_2\theta^{-3} e^{-nm_2\theta^{-2}/2}.$$

So, using Equation (9) to differentiate the product, we have

$$\begin{aligned} L'(\theta) &= C\left\{ \frac{d}{d\theta}\theta^{-2n} \times e^{-nm_2\theta^{-2}/2} + \theta^{-2n} \times \frac{d}{d\theta} e^{-nm_2\theta^{-2}/2} \right\} \\ &= C\left\{ -2n\theta^{-2n-1} \times e^{-nm_2\theta^{-2}/2} + \theta^{-2n} \times nm_2\theta^{-3} e^{-nm_2\theta^{-2}/2} \right\} \\ &= Cn\theta^{-2n-3} e^{-nm_2\theta^{-2}/2} \left(-2\theta^2 + m_2\right). \end{aligned}$$

Since, for $\theta > 0$, $Cn\theta^{-2n-3} e^{-nm_2\theta^{-2}/2} > 0$, $L'(\theta)$ is equal to zero only when

$$-2\theta^2 + m_2 = 0,$$

that is, when $\theta^2 = m_2/2$ or

$$\theta = \sqrt{\frac{m_2}{2}}.$$

(Note that we take the positive square root because $\theta > 0$.)

(c) The maximum likelihood estimator of θ for the Rayleigh distribution is therefore

$$\widehat{\theta} = \sqrt{\frac{M_2}{2}}.$$

Solution to Exercise 7

(a) The maximum likelihood estimate of the parameter p for a geometric distribution is $\widehat{p} = 1/\overline{x}$. For the given data, the sample mean is

$$\overline{x} = \frac{(71 \times 1) + (28 \times 2) + \cdots + (1 \times 6)}{71 + 28 + 5 + 2 + 2 + 1} = \frac{166}{109} \simeq 1.523.$$

Hence the maximum likelihood estimate of p is

$$\widehat{p} = \frac{109}{166} \simeq 0.657.$$

(b) If $Y \sim B(166, p)$, then the maximum likelihood estimate of p is

$$\widehat{p} = \frac{y}{166} = \frac{109}{166} \simeq 0.657.$$

(c) According to Table 11, the estimator in part (a) is biased, while the estimator in part (b) is unbiased. This may seem odd, as the estimator's value is the same in each case. The explanation is that the modelling assumptions differ in the two cases. In part (a) the sampling model is a geometric distribution, while in part (b) it is a binomial distribution.

Acknowledgements

Grateful acknowledgement is made to the following sources:

Page 83 top: © Ajna Rivera

Page 83 bottom: Maggie McCain / https://commons.wikimedia.org/wiki/File:Upper_case_and_lower_case_types.jpg. This file is licensed under the Creative Commons Attribution Licence http://creativecommons.org/licenses/by/3.0/

Page 84 top: Copyright © Mario Sarto. This file is licensed under the Creative Commons Attribution-Share Alike Licence http://creativecommons.org/licenses/by-sa/3.0/

Page 84 bottom: dionisvera/www.123rf.com

Page 87: © iStock.com/MichaelSvoboda

Page 91: © 2017 American Association for the Advancement of Science

Page 92: Taken from: https://twitter.com/FarmingUK

Page 93: © Na2co3/Dreamstime.com

Page 97: © MLE Films Limited

Page 99: © 1998–2012 Yoshiaki Yoneda

Page 100: Peter Byrne/PA Archive/PA Images

Page 101: © Environment Agency. Reproduced by permission

Page 104: © photod/www.istockphoto.com

Page 106: © Maxim Gertsen/www.123rf.com

Page 113: © Stephanie Frey/Dreamstime.com

Page 114: © iStockphoto.com/IPGGutenbergUKLtd

Page 115: Scanrail/www.123rf.com

Page 119: Allie

Page 123: © iStockphoto.com/franckreporter

Every effort has been made to contact copyright holders. If any have been inadvertently overlooked, the publishers will be pleased to make the necessary arrangements at the first opportunity.